Universidad Politécnica de Madrid
Escuela Técnica Superior de Ingenieros Informáticos

Tesis de máster
MÁSTER UNIVERSITARIO EN INTELIGENCIA ARTIFICIAL

RECONOCIMIENTO DE ENFERMEDADES EN FICHAS TÉCNICAS DE MEDICAMENTOS Y SU ANOTACIÓN CON SNOMED-CT

Autor: Pablo Calleja Ibáñez
Directores: Raúl García Castro,
Asunción Gómez Pérez

Abril 2015

AGRADECIMIENTOS

Me gustaría agradecer a Raúl García Castro por su dirección y guía durante el desarrollo de la tesis. Me gustaría agradecer a Guadalupe Aguado de Cea por los conocimientos aportados en el ámbito lingüístico y me gustaría agradecer a Asunción Gómez-Pérez por su dirección y por darme la posibilidad de realizar mi tesis en el Ontology Engineering Group.

También me gustaría dar las gracias a la Agencia Española de Medicamentos y Productos Farmacéuticos por darnos la posibilidad de realizar el proyecto. Por último, me gustaría dar las gracias a mi familia por el apoyo durante todos estos años.

RESUMEN

La interoperabilidad o habilidad para intercambiar información entre sistemas informáticos es una cuestión de gran importancia en la informática médica. La interoperabilidad influye directamente en la calidad de los sistemas médicos existentes en la práctica clínica, ya que permite que la información se trate de manera eficiente y consistente. Para la comunicación entre sistemas informáticos heterogéneos se necesitan terminologías o diccionarios que representen e identifiquen conceptos médicos de forma única, sin importar el idioma o la forma lingüística en la que aparezcan. Estas terminologías permiten a los sistemas informáticos tener la misma visión del mundo y que la información intercambiada sea entendible.

Actualmente, los esfuerzos para la adopción de estas terminologías en la práctica clínica recaen en los profesionales del dominio médico. Los profesionales son los encargados de reconocer conceptos médicos manualmente en documentos del área de la medicina y anotarlos con el código del concepto asociado en la terminología. No existe ningún método automático que permita el reconocimiento de conceptos de un determinado dominio, como por ejemplo las enfermedades, y que posteriormente encuentre el concepto asociado dentro de una terminología con un grado de precisión suficientemente elevado para que pueda ser adoptado en la práctica clínica.

En esta tesis de máster se propone un nuevo método para el reconocimiento de enfermedades en fichas técnicas de medicamentos y su posterior mapeo con la terminología médica SNOMED-CT en español. El método utiliza dos nuevas técnicas propuestas en la tesis para cada fase. La nueva técnica para el reconocimiento de enfermedades propuesta está basada en reglas y en diccionarios especializados en medicina. La nueva técnica de mapeo está basada en la generación de las posibles combinaciones lingüísticas en las que puede aparecer la enfermedad para realizar comparaciones exactas de palabras, utilizando las funciones sintácticas de las palabras como guía. El método propuesto se centra en la identificación de enfermedades dentro de la sección de indicaciones terapéuticas de las fichas técnicas de medicamentos.

TABLA DE CONTENIDO

Lista de figuras

Lista de tablas

Capítulo 1. INTRODUCCIÓN

1.1. ANTECEDENTES

La informática médica se define como la aplicación de las tecnologías de la información y la comunicación en el área de la medicina (Wyatt & Liu, 2002). El principal objetivo de la informática médica es mejorar la salud general de los pacientes mediante la combinación de conocimientos científicos y de ingeniería básica a problemas relevantes en el ámbito médico.

La informática médica cubre un gran espectro de áreas de investigación como la bioestadística, las historias clínicas digitales, los modelos predictivos clínicos, la percepción de imágenes o la biología computacional (Shortliffe & Cimino, 2006; Wyatt & Liu, 2002). Todas estas áreas de investigación reutilizan recursos y técnicas que se han desarrollado a lo largo de los años en la informática tradicional.

Uno de los mayores problemas que tiene la informática médica es que se sitúa en la intersección entre dos dominios especializados completamente diferentes como son la medicina y la informática. Esto conlleva a que generalmente los expertos de cada dominio no tienen suficientes conocimientos como para investigar o desarrollar algún campo de la informática médica de manera aislada. La informática médica requiere que los expertos de ambos dominios trabajen de manera conjunta.

Generalmente, los ingenieros son los encargados de crear las técnicas y algoritmos para resolver problemas determinados, así como de crear sistemas que puedan ser implantados en la práctica clínica. Por otro lado, el personal médico es el encargado producir los datos que los sistemas utilizan y de revisar los resultados obtenidos por los mismos. En el contexto de la medicina, los errores son de mayor sensibilidad que en otros dominios, por lo que se busca que los sistemas tengan una precisión bastante alta y que cualquier decisión estén supervisada por un humano. Todo este proceso de revisiones continuas por parte de expertos es una tarea lenta y tediosa, lo que dificulta la adopción y usabilidad de los sistemas en la práctica clínica.

1.2. MOTIVACIÓN

Una de las principales áreas de la informática médica es la interoperabilidad (Garde, Knaup, Hovenga, & Heard, 2007; Walker et al., 2005). La interoperabilidad se define como la habilidad de dos o más sistemas o componentes para intercambiar información y utilizarla. En la práctica clínica, requiere que la comunicación entre los sistemas se haga mediante estándares que codifiquen la información de manera única para que el emisor y el receptor de una comunicación puedan entenderse.

Para ello se necesitan terminologías o diccionarios que representen conceptos médicos con identificadores únicos. De esta forma, se consigue que dos sistemas independientes tengan la misma visión del mundo. En la historia de la informática médica se han desarrollado múltiples terminologías que representan el conocimiento médico (SNOMED-CT[1], UMLS[2], MedDRA[3], CIE-10[4], etc.).

En el dominio médico, las fuentes de información abarcan un amplio espectro (pacientes, personal, productos farmacológicos, etc.). La mayoría de estas fuentes de información son producidas por los expertos del dominio de manera manual. Los expertos pueden ser los encargados de introducir manualmente códigos de una terminología o se puede asignar el trabajo a un sistema automático que analice las fuentes de información y codifique. En el último caso, los resultados de la codificación son dependientes de la calidad de la información que produzca el experto en lenguaje natural.

Un paso intermedio dentro del área de la interoperabilidad, es usar como entradas fuentes estandarizadas del dominio médico como las fichas técnicas de medicamentos. Este tipo de fuentes tienen que pasar procesos de validación y certificación para asegurar que la información es completa y correcta. Sin embargo, no tienen un formato común y la información se encuentra en lenguaje natural.

Las agencias reguladoras estatales de medicamentos son las encargadas de suministrar medicamentos a todas las regiones del estado. Debido al gran número de medicamentos existentes, las agencias reguladoras necesitan tener control sobre los medicamentos que están en circulación y cuales se están suministrando para las mismas enfermedades. Por ello, la detección de las enfermedades que tratan dichos medicamentos es de gran importancia.

[1] http://www.ihtsdo.org/snomed-ct
[2] http://www.nlm.nih.gov/research/umls/
[3] http://www.meddra.org/
[4] http://cie10.tiddlyspot.com/

La tarea de identificación automática de enfermedades dentro de textos formales implica toda un área de investigación denominada reconocimiento de entidades nombradas (en inglés *Named Entity Recognition*). Como ya hemos dicho, estas entidades por sí solas carecen de valor para la interoperabilidad del ámbito médico por lo que se intenta asociar dichas entidades con un concepto igual o similar en una terminología médica. Este proceso se conoce como mapeo, y existe un área de investigación dedicada en la que se estudian las técnicas para encontrar conceptos iguales o similares entre terminologías, ontologías o términos en lenguaje natural.

Tras un estudio del estado del arte de las técnicas de reconocimiento de entidades nombradas, se ha observado que no existe ningún método estandarizado para la detección de enfermedades en el ámbito clínico. Las técnicas actuales de mapeo son de propósito general y no están focalizadas para el reconocimiento de entidades que pueden abarcar numerosos descriptores como son las enfermedades. De la misma forma, se ha observado que las técnicas de mapeo actuales no están orientadas al mapeo de conceptos extraídos de textos en lenguaje natural. Las técnicas actuales son muy sensibles a cualquier variación lingüística y no explotan las categorías sintácticas de las palabras que componen los conceptos.

En esta tesis de máster se propone una nueva técnica para el reconocimiento de enfermedades basada en reglas y en diccionarios especializados en medicina. También se propone una nueva técnica de mapeo basada en la generación de las posibles combinaciones lingüísticas en las que puede aparecer la enfermedad. La técnica de mapeo se focaliza en el núcleo del concepto (el nombre). El objetivo del trabajo de la tesis se centra en la creación de un método genérico para el reconocimiento de enfermedades en fichas técnicas de medicamentos y su posterior mapeo con la terminología SNOMED-CT utilizando las dos técnicas anteriores.

El método propuesto en la tesis está orientado al español así como los recursos que utilizan. Parte de estos recursos son generados a partir de un estudio de un conjunto de textos de entrenamiento. El reconocimiento de enfermedades y su posterior mapeo con la terminología SNOMED-CT se centra exclusivamente en la sección de indicaciones terapéuticas de las fichas técnicas de medicamentos.

1.3. ESTRUCTURA DEL DOCUMENTO

El documento está organizado de la siguiente forma:

- El capítulo 2 expone el estado del arte de las técnicas de reconocimiento de entidades nombradas y de las técnicas de mapeo. Después se presenta un apartado para la descripción de los recursos relevantes en el dominio médico utilizados (las fichas técnicas de medicamentos, la terminología médica SNOMED-CT y el diccionario MedDRA).
- El capítulo 3 enumera las deficiencias del estado del arte así como los objetivos y el alcance del trabajo realizado.
- El capítulo 4 plantea el método propuesto para el reconocimiento y posterior mapeo de entidades nombradas en fichas técnicas de medicamentos.
- El capítulo 5 muestra la implementación del método en un sistema para un caso real con fichas técnicas proporcionadas por la Agencia Española de Medicamentos y Productos Sanitarios.
- El capítulo 6 muestra los resultados obtenidos de la implementación.
- El capítulo 7 presenta las conclusiones obtenidas del trabajo.
- El capítulo 8 presenta las líneas futuras a seguir a partir del trabajo.

Capítulo 2. ESTADO DEL ARTE

Este capítulo presenta en primer lugar el estado del arte de las dos principales áreas de investigación que cubre la tesis. A continuación se describen distintos recursos relevantes en el dominio médico que se han utilizado. La sección 2.1 clasifica y describe el estado del arte de las técnicas de extracción de entidades nombradas y la sección 2.2 el estado del arte de las técnicas de mapeos de conceptos a terminologías u ontologías. La sección 2.3 presenta los distintos recursos utilizados. La sección 2.3.1 describe las fichas técnicas de medicamentos. La sección 2.3.2 describe la terminología médica SNOMED-CT. Por último, la sección 2.3.3 describe el diccionario MedDRA.

2.1. RECONOCIMIENTO DE ENTIDADES NOMBRADAS

El reconocimiento de entidades nombradas, en inglés *Named Entity Recogintion* (NER), es una tarea fundamental en los sistemas de extracción de información cuyo objetivo es identificar todas las menciones de entidades nombradas que aparezcan en el texto. Una entidad nombrada se define como un término o frase nominal que identifica a un objeto de un conjunto de otros objetos con atributos similares. El reconocimiento de entidades nombradas se divide en dos tareas: identificación de las entidades y clasificación en conjuntos de categorías. Las tres categorías de entidades nombradas más reconocidas son: personas, localizaciones y organizaciones. Estas categorías se extienden desde conjuntos conocidos como las fechas hasta conjuntos creados específicamente para un dominio particular (ej., enfermedades).

Los sistemas de reconocimiento de entidades nombradas más relevantes pueden ser categorizados por su enfoque en tres grupos: sistemas basados en reglas, sistemas basados en diccionarios (*gazetteers*) y sistemas basados en aprendizaje automático (*machine learning*). Dentro de cada uno de los enfoques existen subcategorías que combinan en diversos grados estas categorías de alto nivel.

Los sistemas basados en reglas utilizan técnicas de identificación de patrones en el texto como heurísticas derivadas tanto de la morfología como de la semántica de las frases de entrada. Generalmente se usan como clasificadores en enfoques de aprendizaje automático (Ikonomakis, 2005; Snow, Jurafsky, & Ng, 2005). Los clasificadores son algoritmos utilizados para asignar un elemento no etiquetado en una categoría concreta conocida de manera automática. Los sistemas basados únicamente en reglas son propensos tanto a saltarse entidades nombradas como a excederse en el reconocimiento de estas.

Los enfoques basados en *gazetteers* utilizan recursos de conocimiento externo para identificar pedazos del texto mediante un diccionario o lexicón construido con nombres de entidades. Los *gazetteers* también proveen de un modelo para la resolución de nombres múltiples para la misma entidad. Este enfoque requiere de un trabajo manual para la creación de los lexicones o de sistemas capaces de crearlos de manera dinámica de alguna fuente o recurso externo. Aunque el proceso de creación puede ser difícil, los sistemas que utilizan este enfoque logran mejores resultados para dominios específicos (Maynard, Tablan, & Ursu, 2001). Muchos de los trabajos relacionados en este área tratan de la expansión de los *gazetteers* a lexicones más dinámicos (Nadeau, Turney, & Matwin, 2006; Toral & Mu, 2006).

Los enfoques estocásticos o de aprendizaje automático son los que mejor resultado obtienen trabajando en diferentes dominios (Baluja, Mittal, & Sukthankar, 2000; L. Zhang, Pan, & Zhang, 2004), y pueden realizar análisis predictivo en entidades que son desconocidas en *gazetteers*. Estos sistemas utilizan modelos estadísticos y algunas formas de identificación de características para hacer predicciones en el texto. Sin embargo, este tipo de enfoques requieren una gran cantidad de datos de entrenamiento anotados para ser efectivos.

Independientemente del enfoque que se utilice, la tarea de reconocimiento de entidades nombradas no es fácil. Esta es una de las tareas de mayor complejidad dentro de los sistemas de extracción de información y sobre la que más trabajos se publican. Los problemas básicos que tiene el reconocimiento de entidades nombradas son (Nadeau & Sekine, 2006):

- Variaciones lingüísticas de las entidades nombradas: los nombres pueden aparecer de diversas formas (ej., John Smith, Mr. Smith, John).
- Ambigüedad de los tipos de las entidades nombradas: el nombre sin contexto puede pertenecer a varios grupos (ej., Washington puede referirse a una ciudad y una persona).

Otros problemas más complejos de la tarea son las cuestiones propias del procesamiento del lenguaje natural, como son la ortografía, la puntuación, el formato, etc.

A continuación se exponen con más detalle las características de los tres grupos de sistemas de extracción de entidades nombradas.

2.1.1. SISTEMAS BASADOS EN REGLAS

La primera tarea al abordar el reconocimiento de entidades nombradas en los sistemas basados en reglas es buscar pistas dentro de la estructura y la gramática del texto que indiquen a los sistemas que las palabras se refieren a alguna entidad

nombrada. Aparte de los problemas del procesamiento del lenguaje natural mencionados anteriormente, la coincidencia de patrones ha resultado tener bastante éxito en la detección de estas entidades, aunque sólo en los corpus formales, como artículos de periódico. Algunos sistemas utilizan la aplicación de reglas como una etapa de preprocesamiento para reducir la complejidad de otras técnicas (Jain et al., 2000).

El estudio de Nadeau (Nadeau & Sekine, 2006) proporciona una tabla de características de nivel de palabra que pueden indicar entidades nombradas. Estas características incluyen diferentes casos: si la palabra es mayúscula o todo en mayúsculas (o caso mixto), puntuación, la morfología y la parte del discurso. Estas características, especialmente capitalización, forman la base para la mayoría de las heurísticas de reconocimiento de entidades nombradas mediante reglas.

Estos indicadores gramaticales pueden extenderse además a una serie de reglas que dependen de patrones posicionales. Por ejemplo, los nombres de pila son generalmente fáciles de identificar por medio de un lexicón de nombres propios, y una palabra en mayúsculas que sigue un nombre de pila es probable que sea un apellido. Otras reglas requieren inspección del contexto que rodea la posible entidad nombrada. En la frase "... comprar 100 acciones de (Palabras +)...", donde (Palabras +) es una serie de palabras en mayúsculas, lo más probable es que sea el nombre de una empresa que cotiza públicamente. Además, preposiciones y otras partes de la oración también pueden ayudar a identificar lugares (ej., Frederick puede ser un nombre o un lugar, pero el uso de la frase "en Frederick" indicaría una ubicación).

Las reglas son recursos que recuperan gran cantidad de entidades nombradas pero necesitan de un estudio del dominio del texto y de una creación de las mismas, generalmente por parte de expertos. Esto hace que las reglas sean muy específicas del contexto en el que se trabaja y que no puedan ser exportadas a otras áreas con la misma efectividad.

Al mismo tiempo, las reglas son dependientes de cada idioma ya que la construcción de oraciones y la ordenación de las palabras son diferentes en cada lengua.

2.1.2. SISTEMAS BASADOS EN GAZETTEER

Los *gazetteers* o diccionarios de entidades se usan como parte fundamental en la extracción de entidades. Estos diccionarios están compuestos por listas de términos propios de un mismo dominio (ciudades, países, títulos, etc.) y permiten la extracción del mismo dentro de un texto.

Los *gazetteers* son utilizados para suministrar conocimiento en el aprendizaje automático, proporcionando candidatos en palabras no entrenadas para su posterior clasificación. También se usan como listas de términos que identificar directamente sobre los textos.

Aunque existen sistemas basados solo en *gazetteers*, suelen usarse de manera conjunta con otros sistemas de reconocimiento de entidades nombradas para ampliar la precisión. Existen sistemas que usan mecanismos basados en reglas, analizadores sintácticos y análisis de frecuencias para proponer candidatos sin enfoques de aprendizaje automático (Kazama & Torisawa, 2007; Nadeau & Sekine, 2006; Richman & Schone, 2008).

Los problemas que acarrean estos recursos son su mantenimiento y escalabilidad, ya que tan solo se limitan a extraer los términos en el texto que sean iguales en el diccionario, por lo que son sensibles a las variaciones del lenguaje. Otro problema añadido es que no resuelve la ambigüedad de la entidad nombrada detectada.

2.1.3. SISTEMAS BASADOS EN APRENDIZAJE AUTOMÁTICO

Los sistemas basados en aprendizaje automático son aquellos que aprenden y clasifican de forma autónoma, mejorando con el paso del tiempo. El principal inconveniente de estos sistemas es preparar el conjunto de entrenamiento inicial para el aprendizaje del sistema. Este conjunto de entrenamiento es un grupo significativo de datos anotados por parte de expertos indicando lo que es o no relevante del texto. El recurso más utilizado de este área son los *conditional random fields* (CRF) y los modelos de Markov.

Conditional random fields

Los CRFs son un tipo de modelo estadístico que se aplica comúnmente en reconocimiento de patrones y en aprendizaje automático, donde se usan para realizar predicciones estructuradas. La diferencia con un clasificador común es que tiene en cuenta el contexto de las muestras vecinas para la predicción. En el procesamiento del lenguaje natural, los CRFs predicen secuencias de etiquetas de clasificación para secuencias de muestras de entrada.

Las muestras de entrenamiento contienen un conjunto de observaciones y etiquetas asociadas a esas observaciones. En el entrenamiento del modelo se asignan unos pesos a cada una de esas etiquetas, indicando su relativa importancia según el caso.

Modelos de Markov ocultos

Los modelos de Markov ocultos se usan normalmente en reconocimiento de entidades nombradas porque la función de etiquetar entidades nombradas tiene una gran relación con la función de etiquetado de los anotadores sintácticos. En

ambos casos, el etiquetado involucra una representación inherente de la secuencia en la que se presenta la oración.

El etiquetado por modelos de Markov ocultos genera etiquetas de entidades nombradas sobre el texto original calculando la probabilidad de que una palabra sea entidad usando frecuencias de n-gramas de un conjunto de entrenamiento. Este método está fuertemente ligado al conjunto de entrenamiento para adquirir resultados correctos.

2.1.4. SISTEMAS NER UTILIZADOS EN EL DOMINIO MÉDICO

En el dominio médico, los sistemas NER se denominan *Medical Entity Recognitio*n (MER). Estos sistemas tratan de detectar y delimitar información referente a entidades médicas en textos y clasificarla en una categoría determinada. Uno de los mayores obstáculos en la tarea de reconocimiento de entidades médicas es la gran variación terminológica que existe en el dominio médico (ej., *Diabetes mellitus type 1*, *Type 1 diabetes*, IDDM o *juvenile diabetes* representan el mismo concepto). Otro problema añadido es la aparición de nuevos nombres o abreviaciones para enfermedades y medicamentos. Estos obstáculos hacen que los enfoques basados en *gazetteers* sean insuficientes en la práctica clínica. La mayor parte de los sistemas MER usan técnicas de aprendizaje automático (CRFs) utilizando como entrada un conjunto de datos anotados. (Bodnari, Del, & Lavergne, 2013; Han & Ruonan, 2011; S. Zhang & Elhadad, 2013). En algunos casos, los CRFs se apoyan en reglas para el reconocimiento de entidades (Abacha & Zweigenbaum, 2011; S. Zhang & Elhadad, 2013).

2.2. TÉCNICAS DE MAPEO

El mapeo se define como el proceso de identificación de un concepto igual o similar a otro. El proceso de mapeo ha sido estudiado por diferentes disciplinas como la recuperación de información, alineamiento ontológico y procesamiento del lenguaje natural para la encontrar mapeos entre conceptos de diversas terminologías y ontologías y entre términos en lenguaje natural y conceptos (Choi, Song, & Han, 2006; Kashyap & Sheth, 1996; Lei Zeng & Mai Chan, 2004; Shvaiko & Euzenat, 2013; Sun & Sun, 2006).

A continuación se exponen las principales técnicas de mapeo orientadas a encontrar correspondencias entre conceptos de acuerdo a J. Euzenat y P. Shvaiko (Euzenat & Shvaiko, 2007). Estas técnicas están orientadas a encontrar de forma automática la correspondencia entre conceptos y se clasifican en cuatro grupos: técnicas léxicas, lingüísticas, estructurales y de aprendizaje automático.

2.2.1. TÉCNICAS LÉXICAS

Las técnicas léxicas son las más básicas. Utilizan los nombres o los términos de los conceptos o entidades para buscar las correspondencias. El principio de la técnica es: cuanto más similares sean los nombres o los términos de dos conceptos, la probabilidad de que el concepto sea el mismo es mayor. Las técnicas léxicas son las que más problemas tienen ya que son sensibles a cualquier variación del lenguaje natural. Los problemas más destacados son:

- **Sinónimos**. Un sinónimo es una palabra o término que tiene el mismo o parecido significado que otra. Sin embargo, no tienen por qué ser lexicalmente similares. En estos casos se necesita de otro tipo de recursos que apoyen la detección ya que las técnicas léxicas no son suficientes.
- **Homónimos**. Un homónimo es cuando una misma palabra es usada para nombrar diferentes conceptos. Por tanto puede darse el caso de que dos conceptos lexicalmente similares denoten conceptos distintos (ej., banco como entidad bancaria o como asiento). El resultado final serán correspondencias incorrectas.
- **Variaciones léxicas**. Al igual que en el reconocimiento de entidades nombradas, trabajar con lenguaje natural implica que existan diferentes formas de presentar un mismo concepto (ej., John Smith, Mr. Smith, John). El uso de abreviaciones y de prefijos o sufijos opcionales también se refleja en este problema (CD, CD-ROM, Disco Compacto).

Dependiendo de cómo se realice la comparación de los términos de las entidades existen diferentes técnicas léxicas. A continuación se exponen con más detalle:

Técnicas léxicas basadas en comparación de cadenas

Las técnicas léxicas de mapeo basadas en comparación de cadenas consideran los términos de los conceptos como secuencias de letras. Actualmente existen muchas aproximaciones de comparación:

- **Comparación exacta**: Se comprueba que los dos términos introducidos son totalmente iguales.
- **N-gramas**: Mide la distancia de igualdad mediante la numeración de bloques secuenciales comunes en ambas cadenas
- **Distancia Hamming**: Esta distancia cuenta el número de posiciones en las que la cadena difiriere.
- **Distancia Levenshtein**. Esta distancia mide el número mínimo de operaciones que habría que realizar para convertir una cadena en otra. Las operaciones incluyen inserción, eliminación y desplazamiento de letras.

Para ayudar a mejorar los resultados de las comparaciones de las cadenas se utilizan normalizaciones léxicas. Esta normalización incluye varios procesos:

eliminación de símbolos y signos de puntuación, eliminación de los espacios en blanco y la conversión de caracteres a minúsculas. En el idioma español también se incluye la eliminación de los acentos.

En español las métricas de aproximación como los N-gramas, la distancia Hamming o la distancia Leveshtein pueden inducir a falsos negativos debido a las variedades léxicas y sintácticas que permite el idioma. En el dominio médico este tipo de métricas no son utilizadas salvo la comparación exacta. Los sistemas clínicos requieren de una fiabilidad muy alta.

Técnicas léxicas basadas en el lenguaje

Las técnicas léxicas de mapeo basadas en lenguaje se basan en el desglose de la cadena por palabras. Estas técnicas se acercan más a las que se usan en el procesamiento del lenguaje natural ya que pueden considerar la estructura gramatical de las cadenas e incluso su significado.

El primer paso que se realiza en este tipo de técnicas es aplicar una normalización lingüística a las cadenas. La normalización suele incluir varios de los siguientes procesos:

- ***Tokenización***: La *tokenización* es el proceso de separar una cadena de texto en unidades de significado mínimo (*tokens*) que son palabras, números y caracteres.
- **Eliminación de palabras vacías**: Las palabras vacías son palabras que aparecen con una frecuencia alta en los textos que no aportan información. Este conjunto de palabras incluyen los artículos, pronombres, preposiciones y conjunciones.
- **Lematización*:*** La lematización es un proceso que consiste en hallar el lema correspondiente de una palabra en una forma flexionada como son el plural, el femenino o la conjugación verbal. El lema es la unidad autónoma del léxico de la palabra y que puede constituir una entrada en el diccionario.

Realizado el proceso de normalización, se pueden aplicar técnicas de comparación de cadenas con las técnicas presentadas previamente o con técnicas basadas en comparación de *tokens*. Las técnicas de comparación de *tokens* miden la similitud de cadenas cuantificando el número de *tokens* comunes en cada una.

Existen técnicas más avanzadas que utilizan a su vez información de la función sintáctica que está realizando una palabra en la oración (Aronson, 2001). Para este tipo de enfoques se necesita un anotador sintáctico propio del idioma en el que se esté trabajando.

El trabajo de Oronoz y colegas (Oronoz, Casillas, Gojenola, & Perez, 2013) utiliza anotaciones para el reconocimiento de entidades en español mediante la

explotación de las funciones sintácticas en la búsqueda de estructuras de causa-efecto. El trabajo se centra en el reconocimiento de enfermedades, medicamentos y sustancias.

2.2.2. TÉCNICAS BASADAS EN RECURSOS LINGÜÍSTICOS

Las técnicas de mapeo basadas en recursos lingüísticos utilizan recursos, normalmente externos, en los que se apoyan para detectar similitudes entre palabras o términos. Típicamente, los recursos externos son diccionarios, tesauros o bases de datos. Estos recursos proporcionan relaciones lingüísticas (sinónimos, hipónimos, etc.) que se utilizan para detectar similitudes o correspondencias entre palabras o términos.

Uno de los recursos más usados en este tipo de técnicas es WordNet. Es una base de datos léxica que agrupa palabras en inglés en conjuntos de sinónimos llamados *synsets*. WordNet proporciona definiciones junto con relaciones semánticas entre los conjuntos de sinónimos (ej., *placental, placental mammal, eutherian, eutherian mammal*).

Eurowordnet es una traducción de Wordnet a varios idiomas europeos, entre ellos el español. Este tipo de recursos tratan de ser genéricos, por lo que no existen recursos especializados en el dominio médico.

2.2.3. TÉCNICAS ESTRUCTURALES

Las técnicas de mapeo estructurales son usadas tradicionalmente en procesos de alineamiento entre ontologías o terminologías. Este tipo de técnicas explotan las propiedades estructurales de las ontologías como las relaciones semánticas para tener información del significado de los conceptos. Las técnicas estructurales basadas en grafos miden la similitud entre dos entidades (procedentes de dos ontologías distintas) en base a la similitud de sus relaciones y vecinos. Para ello se ha de extraer los conceptos vecinos de cada entidad y aplicar métricas entre los vecinos de ambas entidades, generalmente léxicas. Las relaciones jerárquicas son el tipo de relación más usadas en este tipo de técnicas ya que el significado es el mismo tanto en ontologías como en terminologías. El principio básico es comparar las relaciones que tiene los conceptos con sus vecinos. Cuanto más iguales sean esas relaciones mayor similitud existirá.

2.2.4. TÉCNICAS DE APRENDIZAJE AUTOMÁTICO

Las técnicas de mapeo basadas en aprendizaje automático aplican una etapa previa de entrenamiento en la que se muestran las puntuaciones de diferentes técnicas de mapeo sobre sus resultados positivos y negativos. El conjunto de entrenamiento se usa como aprendizaje para clasificar los nuevos mapeos.

Las técnicas de aprendizaje automático resultan útiles en procesos de mapeo donde se aplican múltiples métricas de similitud entre los conceptos. Durante el entrenamiento, se analizan las características de los mapeos positivos y negativos para aprender cuales son las más relevantes para determinar si existe o no mapeo entre dos conceptos. Este tipo de técnicas permite llevar a cabo con mayor garantía la selección del mapeo más adecuado entre un conjunto de candidatos en situaciones con la misma puntuación.

2.3. RECURSOS RELEVANTES EN EL DOMINIO MÉDICO

A continuación se presentan los recursos relevantes en el dominio médico que se han utilizado para el desarrollo de la tesis.

2.3.1. FICHAS TÉCNICAS DE MEDICAMENTOS

Cuando un medicamento se comercializa en nuestro país se aprueba su utilización para unas determinadas indicaciones y condiciones de uso. La ficha técnica o resumen de las características del producto es el documento oficial, aprobado por una agencia reguladora, en el que se recoge la información científica esencial destinada a los profesionales sanitarios sobre diferentes aspectos del medicamento: indicaciones, posología, contraindicaciones, efectos adversos, precauciones para su empleo y condiciones de conservación. Las indicaciones y las condiciones autorizadas de uso de un medicamento se corresponden con las que se han estudiado en la fase de investigación clínica y para las que la agencia reguladora garantiza un balance favorable entre el beneficio y el riesgo poblacional.

La práctica de prescribir medicamentos fuera de las condiciones de uso autorizadas entraña riesgos para el paciente y puede constituir una fuente de litigios para el profesional sanitario, siendo utilizada la ficha técnica, en calidad de documento oficial, como referencia ante los tribunales para valorar la actuación profesional del médico[5,6].

La Ley del Medicamento[7], en su artículo 8 establece que se considera medicamento cualquier sustancia medicinal dotada de propiedades para prevenir, diagnosticar, tratar, aliviar o curar enfermedades. También se consideran como medicamentos las sustancias dotadas de estas propiedades, aunque en su comercialización no se

[5] Ficha técnica: ¿qué es y qué implicaciones tiene? INFAC 2002.
[6] Real Decreto 767/1993, de 21 de mayo, por el que se regula la evaluación, autorización, registro y dispensación de especialidades farmacéuticas y otros medicamentos de uso humano fabricados industrialmente. BOE 1993; Julio
[7] Ley 29/2006, de 26 de julio, de garantías y uso racional de los medicamentos y productos sanitarios.

haga referencia a las mismas. Los medicamentos legalmente reconocidos en nuestro país, son: las especialidades farmacéuticas, las fórmulas magistrales, los preparados o fórmulas oficinales y los medicamentos prefabricados.

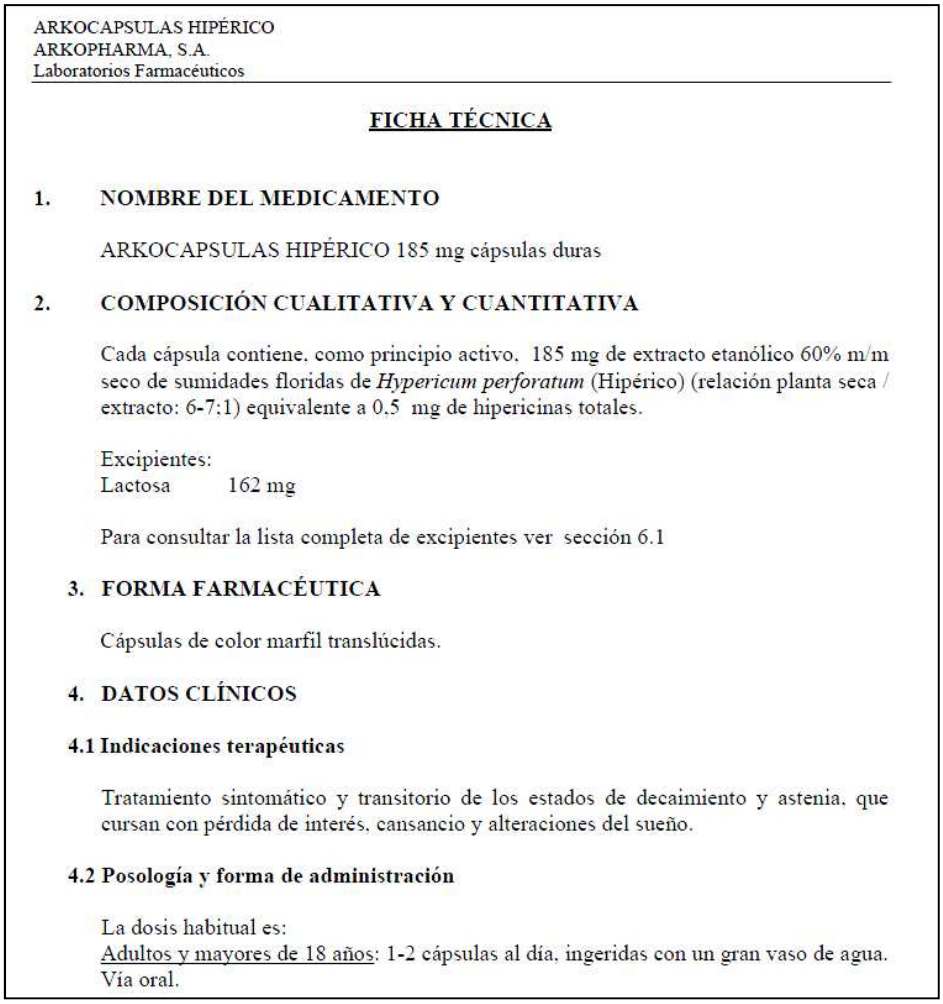
Las agencias reguladoras son los organismos que otorgan la autorización para la comercialización de los medicamentos. En España, dicha autorización depende de la Agencia Española de Medicamentos y Productos Sanitarios (AEMPS) y de la Agencia Europea de Medicamentos (EMA). En el momento en que una agencia reguladora autoriza la comercialización de un medicamento, se aprueban los documentos que garantizan la información disponible sobre el medicamento: la ficha técnica, el prospecto y el etiquetado.

La estructura de la ficha técnica de un medicamento se ajusta a un modelo uniforme y proporciona información actualizada sobre diferentes aspectos del medicamento:

1. Nombre del medicamento
2. Composición cualitativa y cuantitativa
3. Forma farmacéutica
4. Datos clínicos
    - 4.1 Indicaciones terapéuticas
    - 4.2 Posología y forma de administración
    - 4.3 Contraindicaciones
5. Propiedades farmacólogas
6. Datos farmacéuticos
7. Titular de la autorización de la comercialización
8. Números de la autorización de la comercialización
9. Fecha de la primera autorización/ fecha de revisión
10. Fecha de revisión del texto

No es un documento estático, ya que las agencias reguladoras como la AEMPS o la EMA pueden autorizar la modificación del contenido dependiendo de la aparición de nuevas evidencias sobre el medicamento: datos de seguridad a largo plazo, ensayos clínicos para nuevas indicaciones terapéuticas, cambios en los ajustes posológicos, etc. En la Figura 1 se muestra el ejemplo de ficha técnica de medicamento.

El método propuesto en la tesis se centra en la identificación de enfermedades dentro de la sección 4.1 (Indicaciones terapéuticas) de las fichas técnicas de medicamentos. El objetivo es identificar las enfermedades para las que un determinado medicamento ha sido prescrito.



ARKOCAPSULAS HIPÉRICO
ARKOPHARMA, S.A.
Laboratorios Farmacéuticos

FICHA TÉCNICA

1. NOMBRE DEL MEDICAMENTO

ARKOCAPSULAS HIPÉRICO 185 mg cápsulas duras

2. COMPOSICIÓN CUALITATIVA Y CUANTITATIVA

Cada cápsula contiene, como principio activo, 185 mg de extracto etanólico 60% m/m seco de sumidades floridas de *Hypericum perforatum* (Hipérico) (relación planta seca / extracto: 6-7:1) equivalente a 0,5 mg de hipericinas totales.

Excipientes:
Lactosa 162 mg

Para consultar la lista completa de excipientes ver sección 6.1

3. FORMA FARMACÉUTICA

Cápsulas de color marfil translúcidas.

4. DATOS CLÍNICOS

4.1 Indicaciones terapéuticas

Tratamiento sintomático y transitorio de los estados de decaimiento y astenia, que cursan con pérdida de interés, cansancio y alteraciones del sueño.

4.2 Posología y forma de administración

La dosis habitual es:
Adultos y mayores de 18 años: 1-2 cápsulas al día, ingeridas con un gran vaso de agua.
Vía oral.

Figura 1. Ejemplo de ficha técnica

2.3.2. TERMINOLOGÍA SNOMED-CT

SNOMED-CT (*Systematized Nomenclature of Medicine – Clinical Terms*) es la principal terminología médica multilingüe e integral que existe actualmente. SNOMED-CT pertenece y es mantenida y distribuida por la International Health Terminology Standards Development Organisation (IHTSDO). La IHTSDO es una asociación sin ánimo de lucro que pertenece y es mantenida por sus naciones miembros. En 2014 la asociación contaba con 27 países como miembros pertenecientes y cada uno cuenta con una Edición Nacional para utilizar en su territorio. La edición española se distribuye a través del área de descarga de la página del Ministerio de Sanidad, Servicios Sociales e Igualdad (MSSI)[8].

SNOMED-CT es la principal terminología clínica de referencia seleccionada para la Historia Clínica Digital del Sistema Nacional de Salud, lo que supone un primer paso fundamental hacia la interoperabilidad semántica de la información clínica

[8] https://snomed-ct.msssi.es/snomed-ct/solicitudLicencia.do

del Sistema Nacional de Salud. SNOMED-CT es un vocabulario normalizado que permitirá la representación del contenido de los documentos clínicos para su interpretación automática e inequívoca entre sistemas distintos de forma precisa y en diferentes idiomas, facilitando el uso de la información relevante para la toma de decisiones clínicas.

2.3.2.1. Componentes de SNOMED-CT

La terminología SNOMED-CT se constituye principalmente de 3 componentes:

- **Conceptos**: Los conceptos representan ideas o significados únicos organizados en jerarquías.
- **Descripciones**: Las descripciones son los nombres que se le dan a los conceptos de la terminología SNOMED-CT.
- **Relaciones**: Las relaciones permiten la conexión entre conceptos.

Conceptos

Un concepto es un significado clínico identificado mediante un identificador numérico único (*ConceptID*) que nunca se modifica. El identificador otorga una referencia inequívoca al concepto y no proporciona ningún tipo de información sobre el mismo.

Descripciones

Las descripciones de los conceptos son los términos o nombres asignados a un Concepto de SNOMED-CT. En este contexto, término significa una frase utilizada para nombrar un concepto. Un identificador de descripción (*DescritpionID*) único identifica a una descripción. Varias descripciones pueden asociarse con un concepto identificado por un *ConceptID*. Las descripciones proveen la representación legible por humanos de un concepto. Existen tres tipos de descripciones en SNOMED-CT: descripción completa (en inglés *Fully Specified Name*), término preferido (*preferred term*) y sinónimo (*synonym*).

La descripción completa constituye una forma no ambigua de nombrar a un concepto. No siempre representa la frase más utilizada para describir a ese concepto. Cada descripción completa termina con una etiqueta semántica entre paréntesis que expresa la categoría semántica a la que pertenece el concepto (por ejemplo: hallazgo clínico, sustancia, trastorno, etc.).

El término preferido representa la palabra o frase más habitual utilizada para describir el concepto. A diferencia de la descripción completa, los términos preferidos no necesariamente son únicos. Existen casos en los que el término preferido para un concepto también puede ser el sinónimo o el término preferido de otro concepto.

Los sinónimos representan otros términos que se utilizan para describir un concepto.

Relaciones

Las relaciones conectan los conceptos en SNOMED-CT. El proceso de relación involucra 3 conceptos, los dos conceptos relacionados y un tercero que representa el tipo de relación. Existen dos tipos de relaciones importantes en SNOMED-CT: jerárquicas (llamadas de subtipo o IS_A) y lógicas (de atributo).

Las **relaciones jerárquicas** son las relaciones más frecuentes y en las que se basa la arquitectura de SNOMED-CT. Son usadas para asociar un concepto origen a otro más general; el concepto origen tiene un significado clínico más específico. Este tipo de relaciones componen las jerarquías de SNOMED-CT y van desde el concepto más específico hasta el concepto raíz de SNOMED-CT. En la Figura 2 se muestra un ejemplo de la relación jerárquica entre el concepto *Dermatitis de contacto exfoliativa generalizada* y *Dermatitis exfoliativa generalizada*.

El nivel de detalle clínico de los conceptos aumenta con la profundidad de la jerarquía. La Figura 3 muestra la jerarquía de SNOMED-CT desde el concepto *Dermatitis de contacto exfoliativa generalizada* hasta el concepto raíz *SNOMED-CT*.

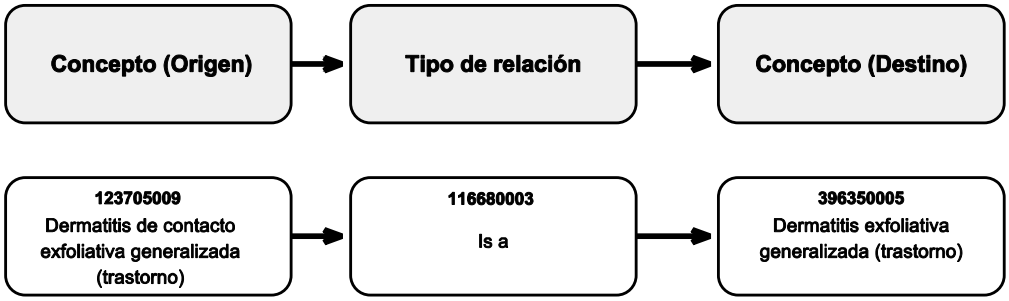


Figura 2. Ejemplo de relación jerárquica en SNOMED-CT

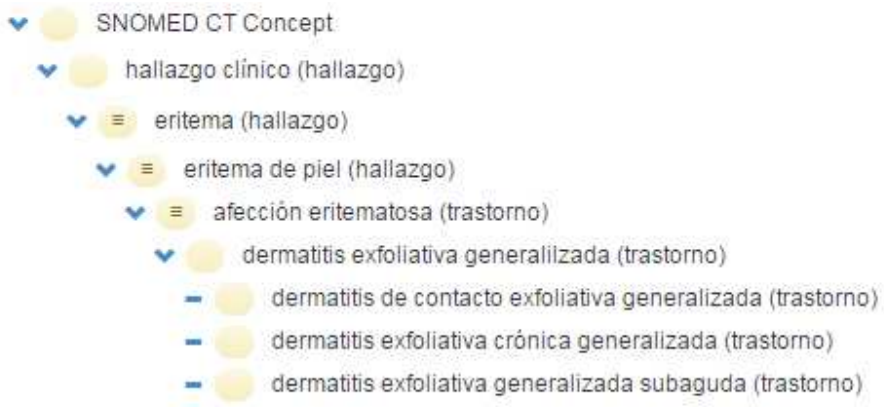


Figura 3. Jerarquía de SNOMED-CT desde el concepto *Dermatitis de contacto exfoliativa generalizada* hasta el concepto raíz *SNOMED-CT*

Las **relaciones de atributo** aportan definiciones lógicas de los conceptos SNOMED-CT, especificando una característica de los conceptos origen de las relaciones. La característica es especificada por el tipo de relación y el valor es definido por el concepto destino de la relación. SNOMED-CT contiene más de 50 tipos de relaciones de atributo. La aplicabilidad de cada tipo de relación está limitada a un determinado dominio y rango. El dominio se refiere a las categorías semánticas válidas para actuar como concepto origen en las relaciones de atributo, mientras que el rango se refiere a las categorías permitidas en el destino de las relaciones.

Los tipos de relaciones de atributo más usadas son: ***interprets*, *procedure site - direct* y *method*.** El tipo *procedure site - direct* describe la parte de cuerpo en la que se realiza un procedimiento clínico, *method* representa la acción que se realiza durante un procedimiento clínico e *interprets* enlaza un hallazgo clínico con la entidad observada. En la Figura 4 se muestra un ejemplo de relación de atributo; para el concepto *Tumor maligno del riñón* se especifica el atributo *sitio de hallazgo* que es el concepto *Estructura del riñón.*

Figura 4. Ejemplo de relación de atributo en SNOMED-CT

Ambiente o localización geográfica	Evento
Calificador	Fuerza física
Componente del modelo de SNOMED-CT	Hallazgo clínico
Concepto especial	Objeto físico
Contexto social	Organismo
Elemento de registro	Procedimiento
Entidad observable	Producto biologico y farmacéutico
Espécimen	Situación con contexto explícito
Estadificaciones y escalas	Sustancia
Estructura corporal	

Tabla 1. Listado de jerarquías de conceptos SNOMED-CT

2.3.2.2. Jerarquía de conceptos

SNOMED-CT tiene 19 jerarquías principales sobre las que cuelgan todos los conceptos de la terminología y todas ellas son nodos hijos del concepto raíz *SNOMED-CT*. Las jerarquías de alto nivel se enumeran en la Tabla 1. La más relevante en el dominio de las enfermedades es *Hallazgo clínico*. Los conceptos de esta jerarquía representan el resultado de una observación, evaluación o juicio clínico e incluyen estados clínicos normales y patológicos.

2.3.2.3. Multilingüismo de SNOMED-CT

Como se ha introducido antes, SNOMED-CT es la principal terminología médica multilingüe que existe actualmente. El proceso de traducción se realiza por las agencias reguladoras de la terminología de cada nación miembro de la IHTSDO. En España, esta agencia es el MSSI que también se encarga de su distribución.

El ciclo de desarrollo de la versión española de SNOMED-CT sigue el mismo ciclo que utiliza la IHTSDO para su actualización y mantenimiento. El ciclo consta de varios procesos: Se obtienen las peticiones de cambio, se edita el modelado y/o las descripciones, se revisa y se valida, se asegura la calidad y finalmente se publica una nueva versión de SNOMED-CT. Cada 6 meses se publica una nueva versión de la terminología en inglés que luego ha de traducirse al español.

El proceso de traducción está reflejado dentro del proceso de edición y/o modelado de las descripciones, y consta de varias tareas. Por cada descripción, se ha de entender el significado representado por la descripción completa (*Fully Specified Name*) en inglés americano, crear una descripción completa en castellano, identificar e incluir sinónimos que representan el mismo concepto y definir uno de los sinónimos como término preferido. Todas las descripciones han de pasar procesos de revisión, control de calidad y validación.

2.3.2.4. Uso de SNOMED-CT

Un estudio de los artículos publicados sobre SNOMED-CT desde el 2001 hasta 2012 revela que el número de artículos incrementa año tras año (D. Lee, de Keizer, Lau, & Cornet, 2014). Sin embargo solo en un 10% del total se exponen el uso o la implementación de SNOMED-CT dentro de la práctica clínica. El porcentaje restante de los artículos se centra en analizar aspectos técnicos de la terminología. Los temas más frecuentes en estas publicaciones son: auditorías de contenidos y estructura de SNOMED-CT (Jiang & Chute, 2009; Wang et al., 2007, 2012), comparación y mapeo de SNOMED-CT con otras terminologías clínicas (Bodenreider & Zhang, 2006; Bodenreider, 2008; Giannangelo & Millar, 2012) y el análisis de la cobertura de SNOMED-CT (De Silva, MacDonald, Paterson, Sikdar, & Cochrane, 2011; Elkin et al., 2006; James & Spackman, 2008).

Debido al gran numero de terminologías médicas fragmentadas con dominios o ámbitos de aplicación solapados, existen bastantes publicaciones sobre comparación y mapeo de SNOMED-CT con otras terminologías. Estos trabajos exponen la necesidad de armonización de las terminologías clínicas en el área médica para la reutilización de la información (Fung & Bodenreider, 2005; Giannangelo & Millar, 2012).

Las publicaciones que analizan la cobertura de SNOMED-CT tratan de medir el grado con el que SNOMED-CT puede codificar correctamente términos locales en contextos específicos. El análisis de cobertura es uno de los primeros requisitos o pasos necesarios para adoptar y utilizar SNOMED-CT en la práctica clínica. Para el trabajo desarrollado en esta tesis, el análisis de la cobertura se focalizó además en la riqueza léxica de la terminología que tiene para cada concepto o jerarquía.

Los estudios que implementan SNOMED-CT en escenarios operativos de la práctica clínica se centran mayoritariamente en desarrollar estrategias para capturar datos de la terminología. La mayoría de estos estudios todavía no han alcanzado la madurez suficiente como para aprovechar los datos capturados. Solo un porcentaje muy pequeño de los artículos presentados usan o experimentan con los datos capturados por SNOMED-CT (Cao et al., 2011; Kim & Park, 2012; N.-J. Lee & Bakken, 2007). Hasta el momento ha habido pocos estudios que aprovechen los datos capturados en SNOMED-CT y, menos aún, que evalúen o cuantifiquen formalmente el valor de SNOMED-CT en escenarios operacionales.

El trabajo desarrollado en esta tesis plantea un nuevo método de extracción y mapeo de enfermedades con SNOMED-CT en un escenario operacional en el que los expertos evalúan el valor obtenido.

2.3.2.5. Mapeo con SNOMED-CT

Gran parte de los trabajos que implementan SNOMED-CT en el ámbito clínico pasan por la búsqueda de correspondencia o mapeos entre términos definidos en lenguaje natural y conceptos SNOMED-CT. Esta búsqueda es un proceso complejo debido al tamaño y granularidad de SNOMED-CT.

Los mapeos pueden dividirse en tres categorías en función del tipo o la procedencia de los datos a mapear: información textual, modelos de datos clínicos estructurados y terminologías clínicas.

En esta sección se exponen varios trabajos y herramientas que han abordado el mapeo de información textual a conceptos de SNOMED-CT. En esta categoría se distinguen dos tipos de información textual: textos clínicos y términos clínicos. Ambos contienen información en lenguaje natural pero se diferencian en la extensión. Un texto clínico puede ser un informe médico o las indicaciones terapéuticas de un medicamento, mientras que un término clínico puede ser una

frase o un concepto escrito de manera aislada. Los diccionarios, glosarios y los registros médicos son fuentes de términos clínicos.

Dentro de los trabajos de mapeo de conceptos SNOMED-CT con textos clínicos existen varios sistemas y herramientas que han demostrado tener éxito en la anotación automática en inglés (Batool, Khattak, Kim, & Lee, 2013; Elkin et al., 2006; Hina, Atwell, & Johnson, 2010; Patrick, Wang, & Budd, 2007; Ruch, Gobeill, Lovis, & Geissbühler, 2008; Stenzhorn, Pacheco, Nohama, & Schulz, 2009). Existen pocos trabajos para la terminología SNOMED-CT en español y generalmente están focalizados en el mapeo de términos clínicos de observaciones médicas (Allones, Hernández, & Taboada, 2014; Castro, Iglesias, Martínez, & Castaño, 2010).

La herramienta desarrollada por P. Ruch y colegas (Ruch et al., 2008) combina dos módulos: el primero aplica varias estrategias de normalización léxica y expresiones regulares para buscar mapeos, mientras que el segundo usa un motor de recuperación de información genérico basado en estructuras algebraicas.

El sistema propuesto por J. Patrick y colegas (Patrick et al., 2007) identifica automáticamente conceptos médicos de SNOMED-CT a partir de datos de pacientes en tiempo real. El sistema incluye varios módulos: extracción de las frases del texto, normalización léxica y equiparación entre frases y conceptos SNOMED-CT basada en la coincidencia de palabras consecutivas (n-gramas).

2.3.3. DICCIONARIO MEDDRA

El Medical Diccionary for Regulatory Activities (MedDRA) es un diccionario terminológico medico creado por la *International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use* (ICH) para facilitar el intercambio de información a través de la estandarización de conceptos.

La ICH es una iniciativa por parte de agencias reguladoras y organizaciones industriales farmacéuticas para la estandarización y coordinación de los aspectos científicos y técnicos en el proceso de regulación de fármacos. La ICH cuenta con la participación de diferentes países y organizaciones a nivel mundial.

MedDRA tuvo sus orígenes en la farmacovigilancia y ha evolucionando para todas las regulaciones médicas del mundo. MedDRA se encuentra en inglés y está traducido a diez idiomas, entre ellos el español. Cada término del diccionario tiene asociado un código numérico único e igual para todos los idiomas, como en la terminología SNOMED-CT.

La estructura del diccionario es parecida a la de SNOMED-CT ya que está clasificada por jerarquías que especifican el grado de profundidad de los términos. La jerarquía más baja denominada *Lowest Level Terms* (LLTs) recoge más de

70.000 términos que reflejan las observaciones reportadas en la práctica. El siguiente nivel, *Preferred Terms* (PTs), son más de 20.000 conceptos simples para la definición de síntomas, signos, diagnósticos, investigaciones o procedimientos. Los PTs relacionados son agrupados dentro de 1.700 términos de la categoría de alto nivel *High Level Terms* (HLTs) basándose en criterios de anatomía, patología, etiología o función. A su vez, Los HLTs están agrupados en 330 grupos denominados *High Level Group Terms* (HLGTs). Finalmente, los HLGTs se agrupan en una clasificación llamada *System Organ Classes* (SOC) que especifica etiología, manifestación y propósito.

Para el mantenimiento, distribución y soporte de MedDRA, la ICH creó la *Maintenance and Support Services Organization* (MSSO). La MSSO cada año publica una nueva versión del diccionario con nuevas actualizaciones y se encarga de la traducción a varios idiomas, entre ellos el español. La MSSO cuenta con desarrolladores con experiencia en el área médica y con conocimiento multilingüe que se encargan del mantenimiento y las traducciones. Aunque la traducción inicial de MedDRA al español fue realizada por la Agencia Española del Medicamento y Productos Sanitarios. La traducción se realiza de manera manual desde los términos LLTs hasta los SOC, asociando los mismos códigos en todos los idiomas.

3. PLANTEAMIENTO

En este capítulo se expone el enfoque seguido para el desarrollo de la tesis. En la sección 3.1 se exponen las deficiencias del estado del arte, a continuación se presentan los objetivos que se quieren alcanzar para superar esas deficiencias en la sección 3.2 y, por último, se muestra el alcance del trabajo en la sección 3.3.

3.1. DEFICIENCIAS DEL ESTADO DEL ARTE

Las deficiencias del estado del arte encontradas mediante el estudio de las técnicas expuestas en el capítulo anterior son las siguientes:

- **Carencia de trabajos de mapeo con SNOMED-CT para español**. Actualmente existen gran cantidad de trabajos aplicables en el ámbito clínico realizados con la terminología SNOMED-CT en inglés (Meizoso, Allones, Taboada, Martinez, & Tellado, 2011; Patrick et al., 2007) y en otros idiomas (Skeppstedt, Kvist, & Dalianis, 2007). Los trabajos realizados para español plantean enfoques generales de mapeo sin llegar a concretar un área específica de trabajo (Cruanes, Romá-Ferri, & Lloret, 2012).

- **Deficiencia en el reconocimiento de entidades nombradas en el ámbito clínico.** No existe ningún método estandarizado, en inglés o español, para la extracción de entidades del dominio específico de las enfermedades. Trabajos como (Castro et al., 2010; Castro & Martinez, 2007; Cruanes et al., 2012; Meizoso García, Iglesias Allones, Martínez Hernández, & Taboada Iglesias, 2012) identifican de forma automática términos procedentes de observaciones simples. No se extraen de textos formales como fichas técnicas de medicamentos.

- **Deficiencias en el mapeo terminológico**. Las técnicas expuestas presentan una deficiencia a la hora de trabajar con palabras obtenidas del lenguaje natural. No se explotan las categorías sintácticas de las mismas y la posibilidad de mapear conceptos similares respecto al núcleo del término. Cualquier variación lingüística propia del lenguaje natural conduce a falsos positivos que en el ámbito médico no se pueden permitir.

3.2. OBJETIVOS DEL TRABAJO

En base a las deficiencias encontradas en el estado del arte actual, se exponen los objetivos que el trabajo quiere alcanzar:

- **Creación de una técnica de reconocimiento de enfermedades**. En la tesis se propone una técnica de reconocimiento de entidades nombradas para enfermedades basada en reglas y en *gazetteers* apoyados en diccionarios especializados en medicina. Al mismo tiempo, las reglas utilizan *gazetteers* de moderadores lingüísticos para delimitar la detección de entidades nombradas. Esta técnica permite obtener entidades nombradas dentro de textos en lenguaje natural. La tesis se centra en el reconocimiento de enfermedades en fichas técnicas de medicamentos, concretamente en la sección de indicaciones terapéuticas.

  Como las reglas de reconocimiento de entidades nombradas son insuficientes para la detección de todas las enfermedades relevantes, se plantea como hipótesis de investigación:
  - H1: la utilización de *gazetteers* basados en diccionarios especializados de enfermedades ayuda en el proceso de reconocimiento de enfermedades.

- **Creación de una nueva técnica de mapeo**. En la tesis se propone una nueva técnica de mapeo basada en comparaciones léxicas de cadenas exactas. La técnica utiliza un algoritmo que genera las posibles combinaciones de la enfermedad detectada atendiendo a sinónimos y al número gramatical de las palabras. El algoritmo explota las categorías sintácticas de las palabras para dar resultados más aproximados y correctos para el ámbito médico. Las enfermedades extraídas por el proceso de reconocimiento de entidades nombradas se mapean con la terminología SNOMED-CT para obtener la descripción y el código identificador del concepto.

  Se plantean como hipótesis de investigación:
  - H2: la utilización de moderadores lingüísticos en el proceso de reconocimiento de entidades nombradas aumenta la precisión de los resultados obtenidos en el proceso de mapeo de enfermedades.
  - H3: la generación de combinaciones basadas en el número gramatical de las palabras o en recursos externos, como los sinónimos, aumentan la precisión en el proceso de mapeo de enfermedades.

- **Creación de un método genérico para el reconocimiento de enfermedades en fichas técnicas de medicamentos y su mapeo a SNOMED-CT en español**. El principal objetivo de esta tesis es la creación de un método genérico que recoja las dos técnicas previamente citadas. El método se focaliza en las indicaciones terapéuticas de las fichas técnicas de medicamentos y consta de una fase de entrenamiento. La técnica de reconocimiento de enfermedades y la técnica de mapeo utilizan los recursos generados por la fase de entrenamiento (también llamada fase de generación de recursos de conocimiento). El método propuesto en la tesis está orientado para mapeos de conceptos con la terminología SNOMED-CT en español para la que existen muy pocas aplicaciones actuales. Los recursos utilizados en la tesis se encuentran también en español.

3.3. ALCANCE

El alcance de este trabajo es el siguiente:

- El método propuesto en la tesis está orientado al español así como los recursos que utiliza.

- La técnica de reconocimiento de entidades nombradas está basado en reglas y *gazetteers* creados a partir de un estudio de un conjunto de textos de entrenamiento.

- El método de reconocimiento de enfermedades y su mapeo con la terminología SNOMED-CT se centra exclusivamente en la sección de indicaciones terapéuticas de las fichas técnicas de medicamentos.

- La técnica de mapeo utiliza un subconjunto de la terminología SNOMED-CT compuesto por las jerarquías de alto nivel que tienen relación con las enfermedades (ej., Hallazgo clínico).

- La técnica de mapeo utiliza recursos generados por la fase de entrenamiento.

Capítulo 4. MÉTODO PARA EL RECONOCIMIENTO Y MAPEO DE ENFERMEDADES

Este capítulo presenta el método propuesto para el reconocimiento de enfermedades de fichas técnicas de medicamentos y su posterior mapeo a la terminología SNOMED-CT. La sección 4.1 presenta la visión general del método y las fases en las que se divide. La sección 4.2 expone los recursos externos que se utilizan en el método. La sección 4.3 describe la primera fase del método en la que se extrae la información necesaria de las fichas técnicas de medicamentos. La sección 4.4 describe la fase manual de generación de recursos de conocimiento necesarios para las siguientes fases del método. La sección 4.5 trata sobre la fase de reconocimiento de enfermedades en el texto y la sección 4.6 sobre la fase de mapeo de las enfermedades encontradas a la terminología SNOMED-CT.

4.1. VISIÓN GENERAL

El método propuesto en esta tesis se divide en cuatro fases. Cada fase realiza una función específica y consta de uno o varios procesos. Una de las fases corresponde a la generación de los recursos de conocimiento que se utilizan en el método. Esta fase solo se realiza una vez con un conjunto grande de fichas técnicas. Los procesos realizados en la fase de creación de recursos de conocimiento involucran tareas manuales y estudios por parte de expertos en el dominio.

En la Figura 5 se visualiza el diagrama general del método. Las diferentes secciones bordeadas por distintos tipos de líneas muestran las fases del método. Dentro de cada fase se encuentran los procesos (P) que se realizan. Los procesos están representados como óvalos. Los recursos están representados como rectángulos. Las flechas indican si un recurso es la entrada o la salida de un proceso.

La primera fase es la fase de extracción de la información sobre la que el método trabaja. La fase está compuesta por un único proceso que realiza la extracción del corpus (P1) cuyo recurso de entrada son las fichas técnicas. Un corpus es una colección de piezas de lenguaje que son seleccionadas y ordenadas de acuerdo a un criterio lingüístico explícito con el fin de ser usadas como una muestra del lenguaje. En este caso, el corpus son las indicaciones terapéuticas obtenidas a partir de las fichas técnicas. La salida del proceso es el corpus con la información

terapéutica de cada ficha técnica y una unificación del mismo para la fase de generación de conocimiento.

El método propuesto utiliza recursos de conocimiento generados a partir de un estudio inicial de un corpus de entrenamiento. La fase 2 recoge los procesos encargados de la generación estos recursos. En primer lugar se extrae una lista con las palabras con mayor frecuencia de aparición en el corpus de entrenamiento (P2). En segundo lugar, se realiza un análisis del corpus de entrenamiento mediante herramientas lingüísticas (P3) utilizando la lista de palabras más frecuentes como guía. Los recursos resultantes de este proceso son las palabras vacías, los sinónimos, los moderadores lingüísticos y los patrones y antipatrones léxico-sintácticos. Por último, se generan las reglas léxico-sintácticas (P4) mediante los patrones y antipatrones léxico-sintácticos del proceso anterior.

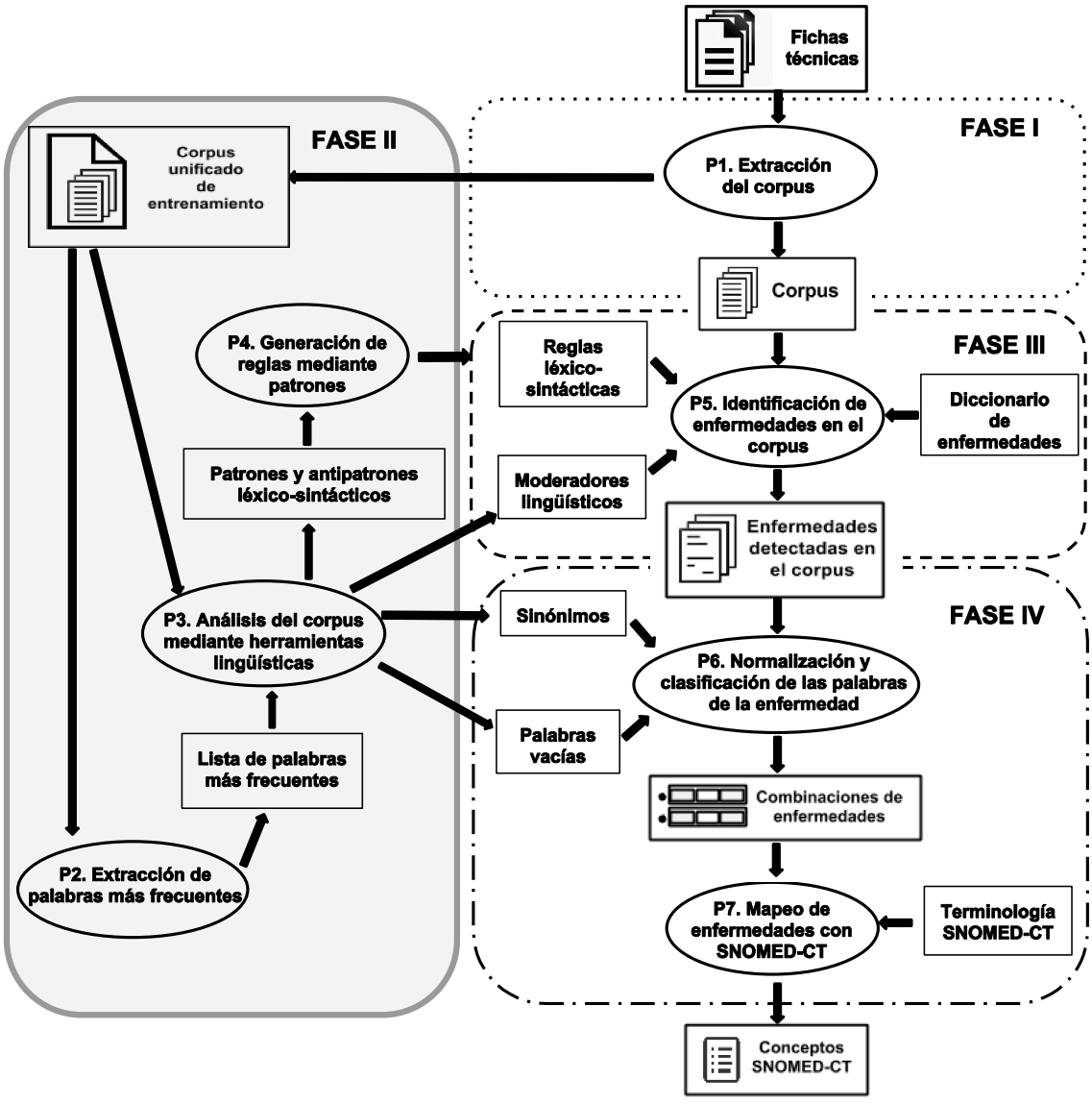


Figura 5. Diagrama general del método propuesto

La fase 3, reconocimiento de enfermedades, consta de un único proceso que tiene como objetivo final la identificación de las enfermedades que aparecen en las fichas técnicas (P5). Estas enfermedades quedan detectadas en el corpus para su posterior mapeo con la terminología SNOMED-CT. El proceso utiliza reglas léxico-sintácticas basadas en los patrones y antipatrones, moderadores lingüísticos y un diccionario de enfermedades para la detección de enfermedades sobre el corpus.

La fase 4 tiene como objetivo el mapeo de las enfermedades encontradas a conceptos equivalentes en la terminología SNOMED-CT. La fase consta de dos procesos. El primer proceso se encarga de normalizar y clasificar las palabras que aparecen en cada enfermedad (P6). El proceso utiliza las listas de sinónimos y palabras vacías para la normalización. Mediante la normalización se crean las posibles combinaciones de la enfermedad mediante los sinónimos y las variaciones de la normalización. El resultado del proceso son las combinaciones de la enfermedad con las palabras normalizadas y clasificadas.

En el proceso de mapeo (P7), se recogen las combinaciones y se aplica el algoritmo de mapeo propuesto en esta tesis con la terminología SNOMED-CT. El resultado del proceso es una lista con los conceptos y códigos identificadores encontrados en la terminología para cada enfermedad que aparece en la ficha técnica.

4.2. RECURSOS EXTERNOS

El método propuesto en la tesis utiliza dos recursos externos que no son generados en la fase 2 de generación de recursos de conocimiento por parte de expertos. El primero es un diccionario de enfermedades sobre el que se apoya el proceso de identificación de enfermedades en el corpus. El segundo recurso es la terminología SNOMED-CT sobre la que se realizan los mapeos.

4.2.1. DICCIONARIO DE ENFERMEDADES

El diccionario de enfermedades es una lista de nombres de enfermedades sin ningún tipo de información adicional. Se requiere que el diccionario tenga un número significante de entradas debido a que el método utiliza este recurso como un *gazetteer* de entidades nombradas de enfermedades.

4.2.2. TERMINOLOGÍA SNOMED-CT

La terminología SNOMED-CT es muy pesada y difícil de manejar ya que cubre todas las áreas médicas. Un objetivo que se plantea es evitar en la medida de lo posible obtener conceptos fuera del dominio de las enfermedades. Por ello, se ha procedido a realizar una poda sobre la terminología de aquellas áreas relacionadas con las enfermedades.

Tras un estudio exhaustivo de los trabajos realizados con la terminología médica SNOMED-CT se ha concluido que no existe ningún método estandarizado para podar conceptos que estén fuera del marco de trabajo. La *International Health Terminology Standards Development Organisation* (IHTSDO) da la posibilidad de crear, distribuir y mantener subconjuntos de conceptos que engloben a un área específica. Pero la metodología no contempla la poda de SNOMED-CT por tipos de conceptos o relaciones; los subconjuntos se hacen con grupos de expertos que seleccionan a mano aquellos términos que consideran relevantes en el dominio. En nuestro caso, el conjunto de enfermedades descritas en SNOMED-CT es demasiado amplio para seguir este método.

Para identificar los conceptos de la terminología que son de interés, se han utilizado como guía los descriptores de grupos. Estos descriptores aparecen entre paréntesis en las descripciones de tipo 3 (o preferidas) y detallan el grupo SNOMED-CT al que pertenece el concepto mediante un término. Se ha verificado que al menos una de las descripciones de cada concepto tiene un descriptor asociado.

Tras estudiar el conjunto de descriptores de grupos de la terminología se han seleccionado aquellos que hacen referencia a las enfermedades. Los descriptores de grupos relacionados son: *hallazgo clínico, trastorno, anomalía morfológica, estructura calcificada y estadificación tumoral*. Estos descriptores de grupos se usan como guía a la hora de podar SNOMED-CT. Si un concepto contiene, en sus múltiples descripciones, el término de uno de estos descriptores, se deja en la terminología. Mediante esta metodología se ha podado SNOMED-CT dejando el subconjunto de enfermedades en aproximadamente 310.000 descripciones de las 1.074.545 incluidas originalmente.

Para la creación del recurso que se utiliza en el método se realiza una normalización sobre las descripciones de los conceptos de la terminología podada. La normalización deja todas las letras en minúsculas, elimina acentos y quita palabras vacías como determinantes, artículos y conjunciones. Estas descripciones normalizadas se utilizan durante el proceso de mapeo.

4.3. FASE DE EXTRACCIÓN DE LA INFORMACIÓN

El objetivo de la fase 1 de extracción de la información es la creación del corpus a partir de las fichas técnicas de medicamentos. En el proceso P1 se extraen las indicaciones terapéuticas de las fichas técnicas de medicamentos. En la Figura 6 se muestra el diagrama del proceso y los resultados. El resultado del proceso será un corpus que se compondrá de ficheros con las indicaciones terapéuticas de cada medicamento. Al mismo tiempo se creará un único fichero que contenga el corpus

unificado para facilitar el estudio y el procesamiento del mismo en la fase 2 de generación de recursos de conocimiento.

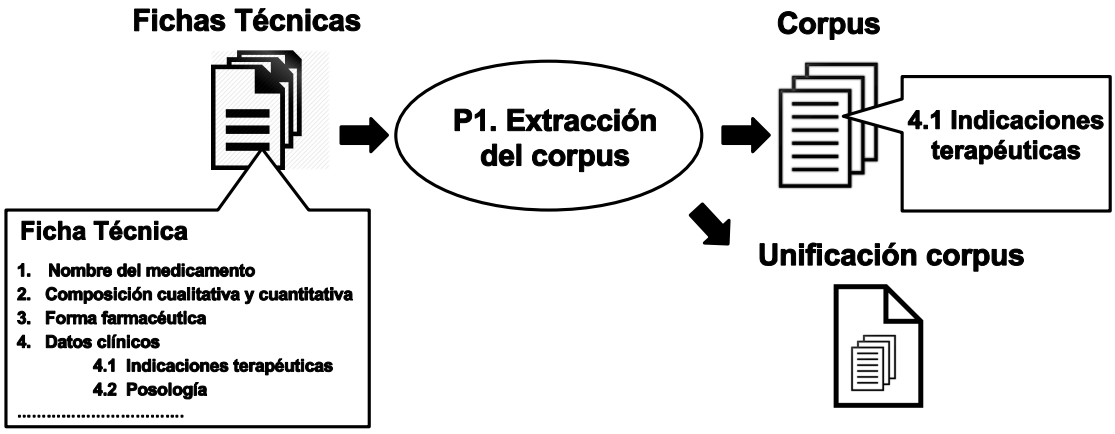


Figura 6. Diagrama del proceso de extracción del corpus

Durante el proceso de extracción del corpus se realizan también dos tareas de limpieza de la información. Primero se añade un punto al final de cada indicación terapéutica en caso de que no termine con un signo de puntuación. La segunda tarea consiste en la eliminación de caracteres especiales. Ambas tareas ayudan al procesamiento del corpus en siguientes fases.

4.4. FASE DE GENERACIÓN DE RECURSOS DE CONOCIMIENTO

Para las fases 3 y 4, reconocimiento de enfermedades y mapeo con la terminología SNOMED-CT, el método propuesto requiere de una serie de recursos creados a partir de un conjunto de procesos que estudian un corpus grande de referencia. Este conjunto de procesos, que involucran tareas manuales, están recogidos en la fase 2 de generación de recursos de conocimiento.

Como se visualiza en la Figura 5, la fase 2 de generación de recursos de conocimiento, está compuesta por tres procesos. En primer lugar se extraen las palabras con mayor frecuencia de aparición (P2). El siguiente proceso (P3) consiste en analizar el corpus mediante herramientas lingüísticas utilizando la lista de palabras del proceso anterior como guía. El resultado es la identificación por parte de expertos de patrones y antipatrones léxico-sintácticos, una lista de moderadores del lenguaje, una lista de sinónimos y una lista de palabras vacías. Para incorporar al método los patrones en lenguaje natural se ha de pasar por un proceso manual (P4) que cree las reglas léxico-sintácticas de extracción de patrones y antipatrones para que trabajen sobre el corpus de manera autónoma.

4.4.1. EXTRACCIÓN DE PALABRAS MÁS FRECUENTES

El objetivo de este proceso es la extracción de forma automática de una lista con las palabras que aparecen en el corpus y su frecuencia de aparición. El resultado del proceso es un fichero CSV en el que la primera columna corresponde a la palabra y la segunda al número de apariciones en el corpus. En la lista no se tienen en cuenta las palabras vacías como artículos, pronombres o preposiciones. Esta lista se encuentra ordenada de mayor a menor frecuencia de aparición.

La lista pasa por una tarea posterior de revisión manual por parte de expertos. El objetivo de la revisión es eliminar palabras que no tienen relación con el ámbito médico o con las enfermedades. Las indicaciones terapéuticas de medicamentos contienen información no relevante que hacen referencia al medicamento y a la forma en la que se suministra (ej. cápsulas, inyectable, etc.).

4.4.2. ANÁLISIS DEL CORPUS MEDIANTE HERRAMIENTAS LINGÜÍSTICAS.

El proceso de análisis del corpus mediante herramientas lingüísticas consta de cuatro tareas independientes manuales en las que se involucra conocimiento por parte de expertos y herramientas capaces de procesamiento de estadísticas del corpus. De cada tarea se genera un recuso diferente que se utilizan en procesos posteriores del método. Los recursos resultantes del proceso son los patrones y antipatrones léxico-sintácticos, los moderadores lingüísticos, los sinónimos y las palabras vacías.

Identificación de patrones y antipatrones léxico-sintácticos

La tarea de identificación de patrones léxico-sintácticos tiene como objetivo la extracción de estructuras léxico-sintácticas que se repitan en el corpus extraído de las fichas técnicas para que faciliten la identificación de contenido semántico relevante para el dominio. En este caso, la tarea se focaliza en aquellas estructuras que tengan relación con enfermedades. Un identificador claro de que puede existir un patrón léxico-sintáctico es la aparición de ciertas combinaciones de palabras que se repitan en numerosas ocasiones dentro de un texto. Para el estudio de patrones en el corpus de las fichas técnicas se utilizará la lista de palabras más frecuentes del proceso anterior como guía.

Mediante herramientas lingüísticas se estudiarán las concordancias o combinaciones de palabras que aparecen de forma conjunta. Es decir, las palabras que aparecen delante y detrás de un término. Estas combinaciones, formadas por las palabras extraídas de la lista con las combinaciones más frecuentes que se dan en estos textos, son las que facilitan la extracción del conocimiento del dominio. También se estudiarán las co-ocurrencias, es decir, no solo nos fijamos en las combinaciones de palabras que van secuencialmente unidas, sino que también se pueden encontrar aquellas apariciones en las que hay alguna variante entre las

palabras que forman la concordancia (ej., *Indicada en el tratamiento de* e *Indicada para el tratamiento de*).

Para cada patrón identificado también se deben tener en cuenta las excepciones. Una excepción es una estructura léxico-sintáctica correcta que no está haciendo referencia a una enfermedad (ej., *indicada para el tratamiento de adultos*). Estas excepciones son únicas en cada patrón debido a las combinaciones en lenguaje natural que se pueden generar por cada estructura léxico-sintáctica.

En el campo de la medicina se emplean patrones y antipatrones (Patrick et al., 2007) para el reconocimiento de entidades nombradas. Un antipatrón es una estructura léxico-sintáctica que coincide con los requisitos que buscamos en el patrón pero, por motivos ajenos al mismo, el resultado final es erróneo. La principal diferencia respecto a las excepciones es que en este caso la información si es una enfermedad, y es el contexto de la oración el que invalida el resultado (ej., *se desaconseja en el tratamiento de la angina*).

En el caso de las indicaciones terapéuticas los antipatrones no son tan necesarios; la mayoría de la información que se presenta en esta sección va destinada a mostrar con qué enfermedades se receta el producto. Sin embargo, algunos patrones aparecen de forma negada. La identificación de un antipatrón permite anotar la oración para no buscar enfermedades en ella.

Los patrones identificados junto con sus excepciones y los antipatrones se utilizan para la creación de reglas que se apliquen dentro del corpus para la extracción automática de enfermedades.

Identificación de moderadores lingüísticos

Todos los trabajos realizados en el ámbito del procesamiento del lenguaje natural tienen que enfrentarse a múltiples problemas. Una ventaja presente en las fichas técnicas utilizadas es que no contienen errores ortográficos. Estas fichas son escritas por compañías farmacéuticas y las agencias reguladoras como la AEMPS se encarga de revisarlas y mantenerlas. Aun así, las descripciones de los medicamentos no están estandarizadas y la riqueza que permite el lenguaje natural es enorme. Por ello, los patrones no son suficientes para determinar dónde se encuentra una enfermedad en el corpus. El principal problema que aparece con la utilización de patrones es determinar dónde parar la búsqueda de la enfermedad. Muchas fichas técnicas disponen de elementos paralingüísticos, como los signos de puntuación, que darán por finalizada la búsqueda. Pero, en muchos otros casos, aparecerán frases subordinadas, conjunciones o descripciones de la enfermedad que introducirán ruido en la detección y en el método.

La tarea de identificación de moderadores lingüísticos se realiza mediante herramientas lingüísticas para estudiar grupos de palabras que acompañan a las

enfermedades y que no producen ningún valor añadido, y que podrían utilizarse como punto de corte. En esta tarea se crean dos listas de moderadores lingüísticos que se recogen en un *gazetteer*.

La primera lista es la que contiene palabras que indican el principio de una frase, subordinadas o no. También se añaden palabras que indican el principio de un patrón en caso de que la búsqueda no haya parado hasta ese momento. La segunda lista es la que contiene expresiones de gradación de enfermedades, es decir, combinaciones de palabras que describen a la enfermedad o a quién va indicado el producto y que produce ruido en el proceso de mapeo con la terminología SNOMED-CT (ej., de moderado a grave, en pacientes adultos, etc.).

Identificación de sinónimos, acrónimos y siglas

La tarea de identificación de sinónimos, acrónimos y siglas se realiza mediante un estudio del corpus e información proporcionada por expertos. Los sinónimos, acrónimos y siglas se agrupan por conjuntos de términos que tienen la misma relación semántica. Los resultados de la tarea de identificación permiten enriquecer el algoritmo de mapeo con la terminología SNOMED-CT. La tarea no se focaliza en cubrir todos los posibles casos, solo los más relevantes.

Identificación de palabras vacías

La tarea de identificación de palabras vacías tiene como objetivo crear una lista de palabras que no aportan información relevante en la fase 4 de mapeo con la terminología SNOMED-CT. La lista se usa en el proceso de normalización y clasificación de las palabras de la enfermedad. Las palabras vacías son en su mayor parte determinantes, artículos, y conjunciones.

4.4.3. CREACIÓN DE REGLAS MEDIANTE PATRONES

El proceso de de creación de reglas léxico-sintácticas mediante patrones tiene como objetivo la implementación de los patrones con sus excepciones y los antipatrones obtenidos en el proceso anterior en reglas que se ejecuten de forma automática. Las reglas obtenidas incluirán los signos de puntuación y las anotaciones para detectar moderadores ligústicos como puntos de corte. Es importante destacar en las reglas qué se considera enfermedad, ya que puede no estar incluido en el mismo (como en el caso de *enfermedad de Crohn*).

4.5. FASE DE RECONOCIMIENTO AUTOMÁTICO DE ENFERMEDADES EN EL CORPUS

La fase 3, reconocimiento de enfermedades, tiene como objetivo identificar las enfermedades que se encuentran dentro del corpus resultante de la fase 1. La fase se realiza mediante un único proceso (P5) que se realiza de forma automática y sin supervisión. El proceso utiliza el diccionario de enfermedades, los moderadores lingüísticos y las reglas léxico-sintácticas.

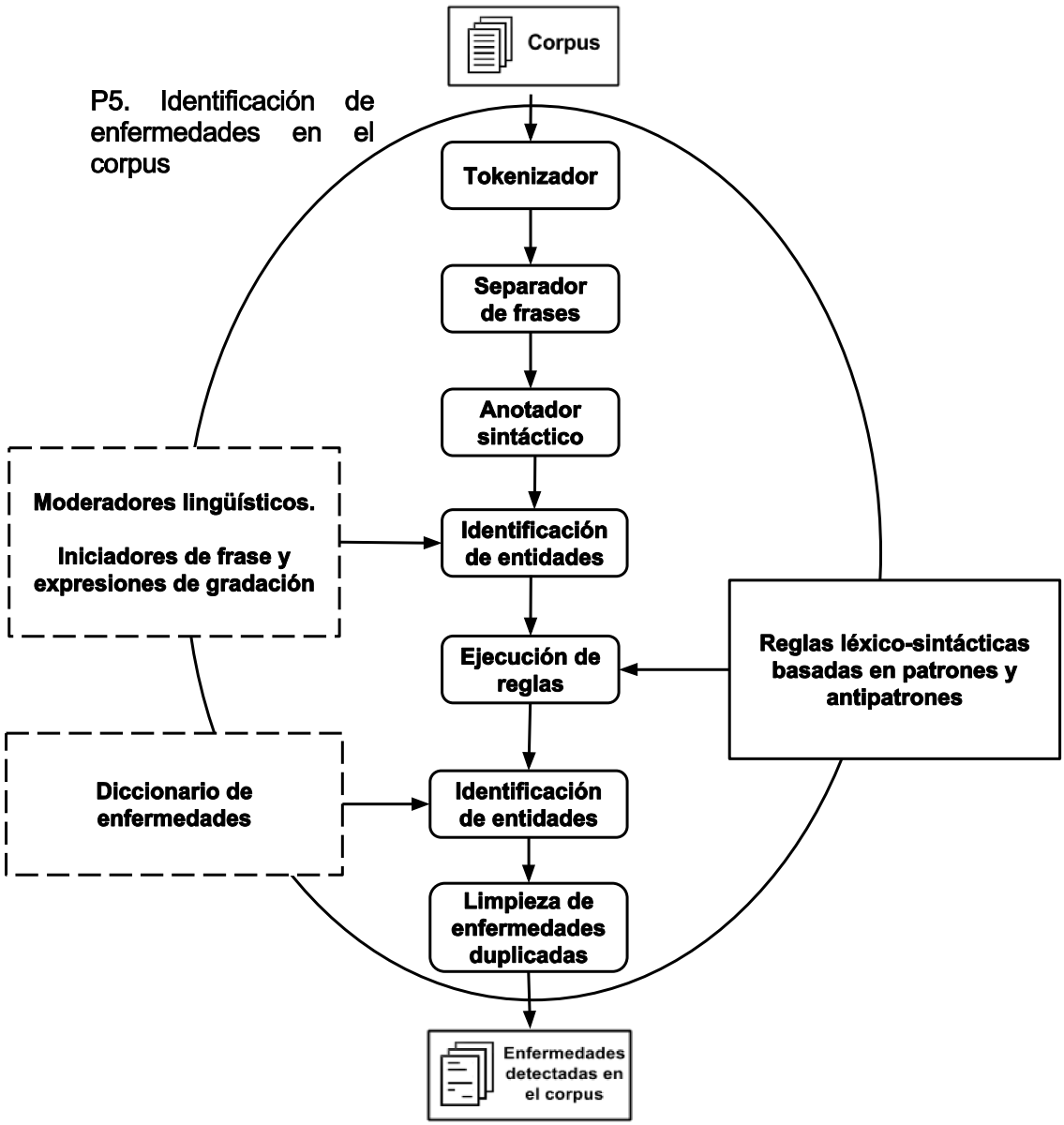


Figura 7. Diagrama de ejecución del proceso de identificación de enfermedades en el corpus

El proceso trabaja como un extractor de información realizando una serie de tareas en un orden determinado. La Figura 7 muestra el diagrama de ejecución de dichas tareas. En primer lugar se realiza un *tokenización* del documento, se reconocen las unidades mínimas de significado y se marca su tipo (palabra, signo de puntuación o número) y su sintaxis (mayúsculas, solo la primera en mayúscula o minúsculas). Después se ejecuta el separador de frases, que se encarga de agrupar los grupos de *tokens* en frases. La siguiente tarea es dotar a los *tokens* de las frases de la funcionalidad sintáctica que ejercen en la frase. Para reconocer la funcionalidad sintáctica de las palabras se utiliza un etiquetador sintáctico en español.

Realizadas estas tareas se procede a la detección de los moderadores lingüísticos. Los moderadores lingüísticos, identificados en la fase 2 de generación de recursos de conocimiento, recogen en un *gazetteer* iniciadores de frase y expresiones de gradación de enfermedades. La tarea de detección de entidades utiliza el *gazetteer* para identificar los moderadores lingüísticos que aparezcan en el corpus. En la Figura 7 los *gazetteers* se muestran en rectángulos con línea discontinua.

La identificación de moderadores lingüísticos sirve como guía en la tarea de ejecución de reglas de reconocimiento de enfermedades basadas en patrones. El conjunto de *tokens* encontrados por el patrón se anotan como enfermedad. Al mismo tiempo se ejecutan los antipatrones, que marcan los *tokens* encontrados por el patrón hasta el final de la frase para evitar la detección de enfermedades en ese segmento del texto.

La siguiente tarea consiste en utilizar el *gazetteer* del diccionario de enfermedades para identificar entidades en el corpus. Las entidades identificadas también se anotan como enfermedad. Por último se ejecuta un modulo de limpieza que elimina anotaciones duplicadas de enfermedades o aquellas que no contienen ni nombres ni adjetivos.

4.6. FASE DE MAPEO

La fase 4 de mapeo tiene como objetivo proporcionar un concepto igual o similar en la terminología SNOMED-CT a una enfermedad dada. La fase consta de dos procesos. El primer proceso se describe en la sección 4.6.1 y se encarga de normalizar las palabras de la enfermedad y de crear las posibles combinaciones en las que puede aparecer mediante el recurso de palabras vacías y el recurso de sinónimos, acrónimos y siglas. El segundo proceso se describe en la sección 4.6.2 y se encarga del mapeo de las combinaciones de la enfermedad a la terminología SNOMED-CT mediante una nueva técnica propuesta en este trabajo.

4.6.1. NORMALIZACIÓN Y PREPROCESAMIENTO

El proceso de normalización y preprocesamiento permite preparar las enfermedades obtenidas en el proceso anterior para que puedan ser tratadas por el algoritmo de mapeo. Este preprocesamiento consiste en limpiar el texto de entrada y generar las posibles combinaciones en las que la enfermedad puede estar escrita. Todas las palabras se pasan a minúscula, se quitan acentos y se eliminan las palabras vacías mediante el recurso de palabras vacías. Una vez se tiene el texto limpio y sin palabras vacías, las palabras restantes se clasifican en dos grupos principales: palabras clave o normales. En el grupo de las palabras clave van a estar solo los nombres, que serán las claves por las que el algoritmo se guiará. El resto de palabras como verbos, adjetivos, negaciones o números pasan a una categoría descriptora o de segundo plano.

Con las palabras categorizadas se procede a crear el conjunto de combinaciones. La combinación inicial es la composición de los *tokens* identificados como enfermedad en el corpus. Sobre ésta, se comprueba que todas las palabras estén en singular. En caso de que existan palabras en plural, se generará una nueva combinación normalizando el conjunto de palabras que componen la enfermedad. Por último, mediante la lista de sinónimos se comprueban las palabras de las combinaciones. Si alguna de las palabras coincide, se generan nuevas combinaciones por cada sinónimo como en la Figura 8. En la figura se muestran las palabras clave en negrita y subrayadas.

Otra tarea que se realiza en este proceso es la comprobación de palabras compuestas. En el caso de que una palabra tenga un guion en medio, se generan tres combinaciones con la palabra. La primera deja el *token* en su forma original con el guion, la segunda separa las palabras en dos *tokens* y la última junta las dos palabras eliminando el guion (ej., post-parto, postparto, post parto).

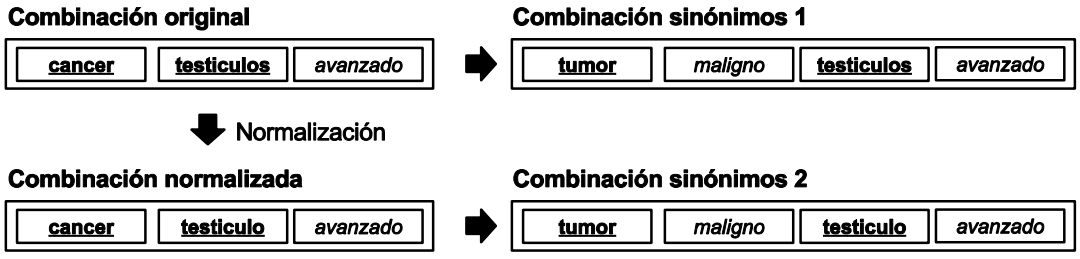


Figura 8. Ejemplos de combinaciones

4.6.2. MAPEO DE ENFERMEDADES CON SNOMED-CT

El objetivo del proceso es mapear cada enfermedad identificada en el corpus con uno o más conceptos de la terminología SNOMED-CT mediante una nueva técnica. La entrada del proceso son los conjuntos de combinaciones de enfermedades y el resultado es una lista con los conceptos y códigos identificadores encontrados en la terminología para cada enfermedad. La técnica se divide en dos fases por las que pasa cada conjunto de combinaciones de una enfermedad.

Fase de búsqueda

En la fase de búsqueda se realiza el algoritmo de mapeo para tratar de encontrar los conceptos candidatos de SNOMED-CT que más se parezcan a las combinaciones de la enfermedad. El algoritmo recoge las combinaciones de la enfermedad y tiene que recorrer todo el recurso SNOMED-CT por cada una. La base del algoritmo es utilizar un umbral que contabiliza el número de palabras iguales, mediante comparación exacta, entre una combinación y un concepto SNOMED-CT. El diagrama de flujo del algoritmo está reflejado en la Figura 9. El sistema guarda como candidatos aquellos conceptos que tengan el mismo valor del umbral, descartando a todos aquellos que estén por debajo. Si se supera dicho umbral, se eliminan todos los candidatos anteriores, se actualiza el umbral y se comienza a guardar nuevos candidatos empezando por el que ha producido el cambio.

El umbral contabiliza y distingue entre palabras clave y las normales. Se priorizan los nombres o palabras clave frente a adjetivos o adverbios, por lo que a la hora de procesar un concepto de SNOMED-CT primero se contabilizan los nombres que han sido encontrados en la combinación y en el concepto SNOMED-CT. Si es mayor se actualiza el umbral borrando los candidatos anteriores. En caso de ser igual se pasa a hacer la misma comprobación con las palabras normales. De esta forma, un concepto que tenga dos palabras clave o nombres y ninguna de menor relevancia será mejor que otro concepto que solo tenga un nombre y tres normales o adjetivos en común.

Por ejemplo, para la combinación *cancer testiculos* de la enfermedad *cáncer de testículos*, el algoritmo recorre SNOMED-CT y va guardando como posibles candidatos aquellos conceptos de la terminología que superen el umbral. En este caso, la terminología SNOMED-CT no tiene ninguna descripción que contenga las dos palabras clave *cancer* y *testiculos*, por lo que los candidatos solo contienen una de ellas (ej., *cáncer de piel*, *herida expuesta de los testículos*, etc.). Con otra combinación de la misma enfermedad, *tumor maligno testiculo*, el algoritmo encuentra nuevos candidatos con dos palabras clave, descartando los anteriores candidatos (ej., *tumor benigno de testículo* y *tumor simple de testículo*). Finalmente, el algoritmo descarta los candidatos anteriores al encontrar el concepto *tumor maligno de testículo* que contiene una palabra normal por encima de los anteriores

candidatos. El resultado para la enfermedad *cáncer de testículos* es el concepto SNOMED-CT *tumor maligno de testículo (363449006)*.

Debido a que el algoritmo trabaja a nivel de *token*, la comprobación entre conceptos negados requiere una especificación concreta. Cada concepto SNOMED-CT se comprueba con la combinación de entrada. Si el concepto SNOMED-CT tiene una negación con la palabra "no" y la combinación también, se comprueba que la siguiente palabra sea iguales en ambos. Si son diferentes el concepto SNOMED-CT se descarta (ej., *cistitis no complicada* y *cistitis no infecciosa*). En el caso de que el concepto no tenga negación y la combinación sí, se comprueba que la siguiente palabra al "no" no aparezca en el concepto (*dolor no cónico* y *dolor crónico*). En caso afirmativo se descarta como posible candidato.

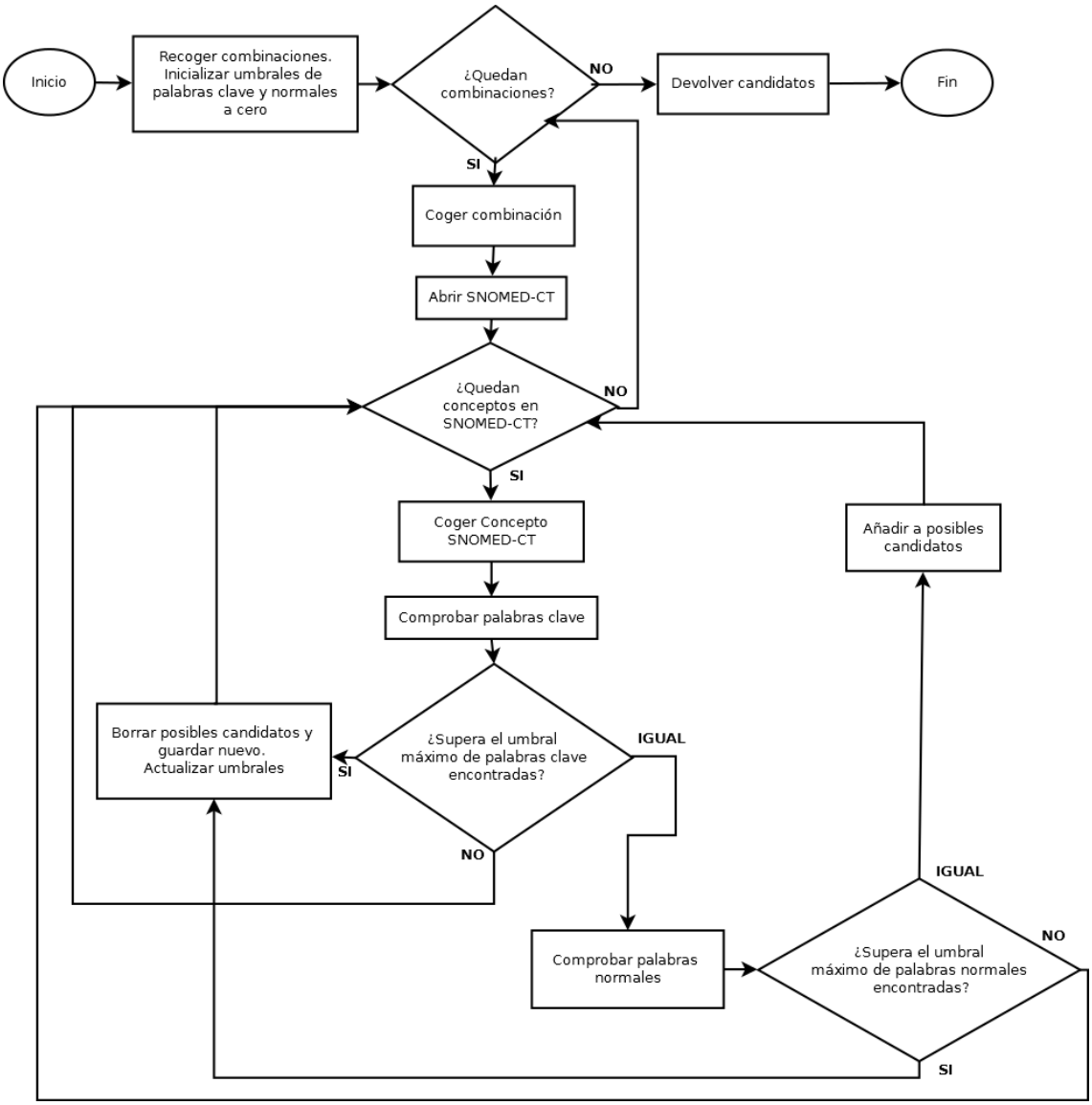


Figura 9. Algoritmo de mapeo de enfermedades para la fase de búsqueda

Fase de selección

La fase de selección está dedicada a la limpieza, organización y selección de los candidatos más óptimos obtenidos por el algoritmo de mapeo. SNOMED-CT es una terminología muy grande y con una amplia granularidad de conceptos por lo que es preciso tener una medida para seleccionar, finalmente, un concepto frente a otro. La medida que se incorpora a cada candidato es la distancia de aproximación al concepto. El cálculo se hace mediante la resta del número de palabras que contiene el concepto SNOMED-CT (sin contar las vacías) y de la suma del total de palabras encontradas, tanto relevantes como no. Esta medida da una cuantificación de la aproximación del concepto frente a la enfermedad detectada siendo mejor cuanto más bajo sea el valor.

Los conceptos candidatos obtenidos por el algoritmo se organizan de menor a mayor por su valor de distancia. Siempre se selecciona el concepto o el grupo de conceptos que tengan el valor más bajo. Cuando existe más de uno en el grupo resultante, se hace una segunda organización dejando en primer lugar aquellos conceptos que tengan adjetivos no específicos como *aguda*.

Siguiendo el ejemplo anterior del *cáncer de testículos,* los conceptos candidatos resultantes por el algoritmo son aquellos que contienen las palabras “tumor”, “maligno” y “testículo”. El candidato óptimo es *tumor maligno de testículo (363449006)* pero el algoritmo también devuelve *tumor maligno de testículo de células de Leydig (278055006).* En el primer caso, la distancia es cero ya que el término se compone de tres palabras (sin contar las vacías) y hay tres palabras encontradas. En el segundo caso, la distancia es dos, el término tiene cinco palabras y tres encontradas. El candidato óptimo es aquel con la distancia más baja: *tumor maligno de testículo (363449006).*

En muchas ocasiones, el sistema devuelve más de tres conceptos SNOMED-CT con la misma distancia, por lo que se concluye que el algoritmo no ha sido lo suficientemente preciso como para concretar un resultado. En estos casos, se amplía el conjunto de palabras poco relevantes extrayendo todos los adjetivos que están presentes en la frase donde se detectó la enfermedad. Con este conjunto, se realiza una segunda búsqueda entre los conceptos candidatos en busca de un óptimo. Por último, el método devuelve los conceptos resultantes obtenidos con su descripción y su código identificador para cada enfermedad.

Capítulo 5. INSTANCIACIÓN DEL MÉTODO

Para demostrar las contribuciones del trabajo se ha instanciado el método propuesto de reconocimiento de enfermedades y mapeo con SNOMED-CT para fichas técnicas proporcionadas por la AEMPS. La instanciación del método se ha realizado mediante la implementación de un sistema que realiza los procesos del método. La implementación del sistema y de los procesos se describe en este capítulo. La sección 5.1 describe los recursos externos que se han utilizado. La sección 5.2 describe la implementación de la fase de extracción. La sección 5.3 explica la fase de generación de recursos de conocimiento y se muestran los resultados obtenidos en cada proceso y tarea. La sección 5.4 describe la implementación de la fase de reconocimiento de enfermedades y la sección 5.5 la fase de mapeo con la terminología SNOMED-CT. Por último, en la sección 5.6 se describe el proceso de publicación de los resultados obtenidos.

5.1. RECURSOS EXTERNOS UTILIZADOS EN EL SISTEMA

La implementación del recurso de la terminología SNOMED-CT para el sistema sigue las directrices propuestas en la sección 4.2.2. Además, la terminología podada y procesada se divide en cuatro ficheros iguales para utilizarse en el proceso de mapeo de manera más eficiente.

El diccionario de enfermedades se genera a partir del diccionario MedDRA. Primero se extraen los *Lowest Level Terms* del grupo de enfermedades del diccionario MedDRA y se eliminan sus códigos de referencia. Por último, se crean dos términos por cada concepto extraído. Uno con la primera letra del concepto en mayúscula y otro en minúscula.

5.2. IMPLEMENTACIÓN DE LA FASE DE EXTRACCIÓN

La implementación de la fase 1 de extracción se realiza con el objetivo de extraer el corpus de las fichas técnicas proporcionadas por la AEMPS. Estas fichas proporcionadas se encuentran divididas por ficheros que corresponden a cada apartado del medicamento. La enumeración de los ficheros no coincide con la organización de apartados de las fichas técnicas vista en el apartado 2.3.1. La AEMPS utiliza una enumeración propia en sus sistemas. Las indicaciones terapéuticas se recogen en los ficheros con el indicador numérico 8. Cada apartado corresponde a un fichero en formato OpenXML y a otro en HTML, donde se

encuentra la información que se desea extraer. El corpus resultante se compondrá de ficheros en texto plano con nombre del identificador de la Base de Datos de la AEMPS y únicamente con la información de las indicaciones terapéuticas.

5.3. IMPLEMENTACIÓN DE LA FASE DE GENERACIÓN DE RECURSOS DE CONOCIMIENTO

Para la fase 2 de generación de recursos de conocimiento se ha utilizado un corpus de 1.078 fichas técnicas proporcionadas por la AEMPS. Los resultados obtenidos por cada proceso son los siguientes.

5.3.1. PALABRAS MÁS FRECUENTES OBTENIDAS

En el proceso de extracción de palabras más frecuentes se ha obtenido una lista de 114 palabras con más de 80 apariciones. De este conjunto, se han eliminado las palabras vacías como determinantes, artículos y conjunciones. Por último, se ha revisado la lista de palabras junto a expertos de la AEMPS y se han descartado aquellas que no tienen relación con las enfermedades. La lista resultante está compuesta por 80 palabras reflejadas en el Apéndice A. En la Tabla 2 se muestran las 6 palabras con mayor frecuencia de aparición sobre las que se ha focalizado el estudio inicial de patrones.

Palabra	Frecuencia
TRATAMIENTO	2130
PACIENTES	984
ADULTOS	701
INDICADO	647
ENFERMEDAD	433
TRASTORNO	412

Tabla 2. Los seis términos más frecuentes de la lista de palabras

5.3.2. ESTUDIO DEL CORPUS MEDIANTE HERRAMIENTAS LINGÜÍSTICAS

A continuación se muestran los resultados obtenidos en cada tarea del proceso de estudio del corpus mediante herramientas lingüísticas.

Patrones léxico-sintácticos identificados

Los patrones léxico-sintácticos identificados se muestran en el Apéndice B. Estos patrones son el resultado de un exhaustivo análisis del corpus mediante herramientas lingüísticas y expertos del dominio. En el estudio inicial del análisis léxico-sintáctico se ha utilizado la lista de palabras más frecuentes como guía para la detección de patrones. Analizando las 6 palabras de la Tabla 2, se ha detectado que *adultos* no aporta información relevante para la detección de enfermedades, por lo que se descarta como palabra de referencia.

Mediante las 5 palabras restantes (*tratamiento, pacientes, indicado, enfermedad y trastorno*) se han identificado 6 patrones léxico-sintácticos. La Tabla 3 muestra los patrones y las palabras que lo construyen, ejemplos y el número de apariciones que tienen sobre el corpus. Cuando el patrón se puede construir con varias palabras diferentes, estas palabras aparecen agrupadas por llaves y separadas por barras (ej., {la/las/los}). Cuando una palabra o conjunto de palabras puede aparecer o no, se marca con un signo de interrogación. Los corchetes identifican aquello que se detectaría como enfermedad.

Patrón	Ejemplos	Nº de apariciones
Tratamiento de {la/las/los} + *[enfermedad]*	• Tratamiento de la hipertensión esencial en adultos • Tratamiento de las crisis mioclónicas en adultos y adolescentes • Tratamiento de los episodios maníacos moderados a graves	701
Tratamiento del + *[enfermedad]*	• Tratamiento del síndrome de Menière • Tratamiento del rechazo de trasplantes en pacientes que previamente han recibido otros agentes inmunosupresores	342
Indicado {en/para} el tratamiento {sintomático/tópico}? {de/del} + *[Enfermedad]*	• Indicado en el tratamiento sintomático de la demencia de tipo Alzheimer de leve a moderada. • Indicado para el tratamiento del cáncer de colon rectal • Indicado en el tratamiento tópico del acné vulgar cuando se presentan pápulas y pústulas	161
Pacientes con + *[Enfermedad]*	• Pacientes con insuficiencia cardiaca que han sido estabilizados • Pacientes con aspergilosis invasiva	451
Enfermedad {de}? + *[Enfermedad]*	• Enfermedad de Alzheimer • Enfermedad vascular periférica	248
Trastorno {de}? + *[Enfermedad]*	• Trastorno obsesivo-compulsivo • Trastorno de ansiedad social	131

Tabla 3. Patrones obtenidos con las palabras *tratamiento, pacientes, enfermedad* y *trastorno*

En el caso de la palabra *tratamiento* se han identificado tres tipos de patrones léxico-sintácticos. Dos de ellos utilizan la misma palabra seguida de *de* o *del* para construir el patrón. El tercero, ha sido hallado mediante la co-ocurrencia entre "indicado" y "tratamiento". Este último patrón es más específico ya que es una estructura léxico-sintáctica propia del ámbito médico para indicar la enfermedad

que trata el medicamento. El patrón cuenta además con adjetivos opcionales que definen el tipo de tratamiento.

Por cada patrón se ha de examinar si cuenta con excepciones, estructuras léxico-sintácticas correctas que no están haciendo referencia a una enfermedad. También se ha de revisar si existen antipatrones en los que el patrón aparece de forma negada. Estos antipatrones se marcan por separado para crear las reglas correspondientes. La Tabla 4 muestra las excepciones y los antipatrones detectados para los patrones de la Tabla 3. En caso de que no existan excepciones o antipatrones para el patrón, la casilla se marca con el símbolo "-".

Patrón	Excepciones	Antipatrones
Tratamiento de {la/las/los} + *[enfermedad]*	• Tratamiento de las siguientes infecciones • Tratamiento de sus signos y síntomas	-
Tratamiento del + *[enfermedad]*	-	-
Indicado {en/para} el tratamiento {sintomático/tópico}? {de/del} + *[Enfermedad]*	• Está indicado para el tratamiento de segunda línea	No está indicado {en/para} el tratamiento {sintomático/tópico}? {de/del} + *[Enfermedad]*
Pacientes con + *[Enfermedad]*	• Pacientes con un alto riesgo de recurrencias • Pacientes con riesgo	-
Enfermedad {de}? + *[Enfermedad]*	• Síntomas asociados a la enfermedad o signos de progresión de la enfermedad. • Cuya enfermedad ha progresado	-
Trastorno {de}? + *[Enfermedad]*	-	-

Tabla 4. Excepciones y antipatrones de los patrones obtenidos por las palabras *tratamiento, pacientes, enfermedad* y *trastorno*

Moderadores lingüísticos identificados

Mediante herramientas lingüísticas, se estudian grupos de palabras que acompañan a las enfermedades que no producen ningún valor añadido y que podrían utilizarse como punto de corte. El resultado de este proceso ha sido la creación de dos listas. La primera lista es la Tabla 5 que contiene palabras que indican el principio de una frase, subordinada o no. También se añaden palabras que indican el principio de un patrón en caso de que la búsqueda no haya parado

hasta ese momento. La segunda lista es la Tabla 6 que contiene expresiones de gradación de enfermedades, es decir, combinaciones de palabras que describen a la enfermedad o a quién va indicado el producto y que produce ruido en el proceso de mapeo con la terminología SNOMED-CT (ej., de moderado a grave, en pacientes adultos, etc.)

asociada	cuyo	está indicado	por ejemplo
asociado	cuyas	hayan pasado	provocada por
asociadas	cuyos	inducida por	que
asociados	después	inducidas por	si
causada por	ej	inducidos por	síntomas asociados
como	en combinación	incluso durante	tras haber
con	en el tratamiento	incluyendo	y con
cuando	en estados iniciales	Mediante	
cuya	está indicada	para los cuales	

Tabla 5. Iniciadores de frase

a largo plazo
a corto plazo
a pesar del
a pesar de
adultos
de bajo impacto
de urgencia
de moderada a grave
de moderado a grave
de moderado a intenso
de moderados a graves
de moderada a severa
de grado moderado a severo
de grado medio o alto
de inicio parcial
de intensidad leve o moderada
de intensidad leve o moderado
de intensidad leve a moderado
de intensidad leve a moderada
de intensidad moderada o severa
de intensidad moderada
de leve a moderado
de leve a moderada
de leves a moderadas
de leves a moderados
de leve a moderadamente grave
de otro origen
en pacientes
en personas
en adultos
en niños
en individuos
en jóvenes
en mujeres
en hombres
en fase aguda
en aquellos pacientes
en los varones
grave
niños
jóvenes
moderado a grave
moderada a grave
moderado a intenso
moderados a graves
metastásico o recidivante
leve a moderadamente grave
leve o moderadamente grave
leve a moderada
leve a moderado
leve o moderada
leve o moderado
leve-moderada
leves
localmente avanzado
por traumatismo
relacionados
subagudas y crónicas
superpuestas a
transitorio o permanente

Tabla 6. Gradaciones de enfermedades

Sinónimos, acrónimos y siglas identificados

La Tabla 7 es el resultado de la tarea de identificación de acrónimos y siglas que permitirán al sistema enriquecer el algoritmo de mapeo con la terminología SNOMED-CT. La tabla se ha realizado mediante un estudio del corpus y de información proporcionada por expertos.

Término	Sinónimo 1	Sinónimo 2	Sinónimo 3
Cáncer	Tumor maligno		
Colorectal	Colon y recto		
CPPC	Cáncer de pulmón de célula pequeña		
Déficit	Insuficiencia	Deficiencia	Carencia
ELA	Esclerosis lateral amiotrofia		
ERGE	Enfermedad por reflujo gastroesofágico		
HBP	Hiperplasia benigna de próstata		
HFHo	Hipercolesterolemia familiar homocigotica		
Infecciones criptococales	Criptococosis		
NVPO	Náuseas y vómitos postoperatorios		
Trastorno emocional	Alteración emocional		
VHS	Virus del herpes simple		
Vulvovaginal	Vulva y vagina		

Tabla 7. Tabla de sinónimos

Palabras vacías identificadas

La Tabla 8 muestra la lista de las palabras vacías identificadas que está compuesta en su mayor parte por determinantes, artículos y conjunciones. Estas palabras no aportan información relevante en el proceso de mapeo con SNOMED-CT.

a	el	la	un
con	en	los	una
cuya	esta	las	y
cuyo	estas	o	
de	esto	que	
del	estos	se	

Tabla 8. Lista de palabras vacías

5.3.3. CREACIÓN DE REGLAS

El proceso manual de creación de reglas tiene como objetivo convertir los patrones léxico-sintácticos con sus excepciones en reglas de un sistema de anotación que permitan la extracción automática de enfermedades en el texto. El sistema de anotación con el que se ha utilizado para el reconocimiento de enfermedades en el corpus es GATE. GATE[9] es una herramienta que provee un entorno de desarrollo y un marco de trabajo para tareas de procesamiento de lenguaje natural y extracción

[9] https://gate.ac.uk/

de información en muchos lenguajes y es desarrollada y mantenida por la Universidad de Sheffield desde 1995. Está desarrollado en Java y el código es abierto. GATE cuenta con un potente lenguaje propio denominado JAPE que permite reconocer anotaciones utilizando gramáticas que están constituidas por pares ordenados de patrones y reglas. JAPE aplica estructuras mucho más complejas que las expresiones regulares. Mediante las reglas JAPE se pueden construir los patrones extraídos del proceso anterior y marcar aquello que se considere enfermedad de manera automática.

Regla basada en patrón	Regla sobre el corpus
"Tratamiento de" {"la"/ "las" / "los" } **[PALABRAS+]** *{SP/MODERADOR}*	• Tratamiento de la **hipertensión esencial** *en adultos* • Tratamiento de los **episodios maníacos** *moderados a graves*
"Tratamiento del" **[PALABRAS+]** *{SP/MODERADOR}*	• Tratamiento del **síndrome de Menière**. • Tratamiento del **rechazo de trasplantes** *en pacientes* que previamente han recibido otros agentes inmunosupresores
"Indicado" {"en"/"para"} "el tratamiento" {"sintomático"/ "tópico"}? {"de" / "del"} **[PALABRAS+]** *{SP/MODERADOR}*	• Indicado en el tratamiento sintomático de **la demencia de tipo Alzheimer** *de leve a moderada*. • Indicado en el tratamiento tópico del **acné vulgar** *cuando* se presentan pápulas y pústulas
"Pacientes con" **[PALABRAS+]** *{SP/MODERADOR}*	• Pacientes con **insuficiencia cardiaca** *que* han sido estabilizados • Pacientes con **aspergilosis invasiva**.
["Enfermedad" {"de"}? **PALABRAS+]** *{SP/MODERADOR}*	• **Enfermedad de Alzheimer**. • **Enfermedad vascular periférica**.
["Trastorno" **{"de"}?** **PALABRAS+]** *{SP/MODERADOR}*	• **Trastorno obsesivo-compulsivo**. • **Trastorno de ansiedad social**.

Tabla 9. Reglas en el corpus basadas en patrones léxico-sintácticos

La Tabla 9 muestra la estructura de las reglas creadas a partir de los patrones detectados en la Tabla 3. En la tabla se muestra el resultado que se obtiene en la ejecución de las reglas JAPE sobre el corpus. Las reglas identifican en el corpus las

palabras que se encuentran entrecomilladas y que forman el patrón. Si el patrón tiene varias opciones de palabras, estas se agrupan entre llaves y separadas por barras. Si la palabra o conjunto de palabras puede aparecer o no, se añade el signo de interrogación. PALABRAS+ identifica un conjunto de palabras hasta que la regla encuentre un signo de puntuación (SP) o un moderador lingüístico (MODERADOR). La regla anota como enfermedad el texto que se encuentre entre corchetes y en negrita. En algunos casos el patrón forma parte de la enfermedad.

5.4. IMPLEMENTACIÓN DE LA FASE RECONOCIMIENTO ENFERMEDADES

La implementación de la fase 3 de reconocimiento de enfermedades se ha realizado como se describió en la sección 4.5. El proceso que se encarga de realizar las tareas descritas en la figura Figura 7 se ha desarrollado en la arquitectura de GATE. Los recursos que utiliza el proceso son las reglas de reconocimiento de enfermedades (patrones y antipatrones lingüísticos) en lenguaje JAPE, la lista de moderadores del lenguaje y la lista de enfermedades del diccionario MedDRA. Ambas listas se usan como *gazeetteers* de entidades en GATE. Para el anotador sintáctico en español se ha utilizado Freeling.

5.5. IMPLEMENTACIÓN DE LA FASE DE MAPEO

La implementación de la fase 4 de mapeo se ha desarrollado en Java en módulos o *plugins* adicionales de la arquitectura de GATE. Para el proceso de de normalización y clasificación de las palabras de la enfermedad (P6) se utiliza información del anotador sintáctico del proceso anterior. El anotador sintáctico anota los *tokens* con la función gramatical que ejerce en la oración. Esta información se usa para encontrar las palabras clave o nombres en la enfermedad.

Además, el anotador sintáctico proporciona el lema de cada palabra. Los lemas son útiles en la tarea de creación de combinaciones de palabras que se encuentran en plural y se tienen que pasar a singular. Los lemas solo se utilizan con los nombres ya que con los adjetivos se puede perder el género (ej., el lema del adjetivo *duras* es *duro*).

Para el proceso de mapeo con la terminología SNOMED-CT (P7) se ha modificado la ejecución del algoritmo descrito en la sección 4.6.2. El algoritmo tiene que comprobar todos los conceptos que se encuentran la terminología por cada enfermedad. Aún con la terminología podada se tienen más de 310.000 términos que han de ser procesados uno a uno. Para agilizar la tarea, se pasa a una ejecución multitarea en la que se crean cuatro hilos de ejecución en los que se aplica el

algoritmo. Cada hilo apunta a un segmento diferente de la terminología que se encuentra dividida en cuatro partes. Los hilos de ejecución van guardando locamente sus conceptos candidatos. Cuando se termina el procesamiento, se comparan entre ellos para determinar quien tiene el mejor grupo de candidatos mediante el umbral más alto, descartando el resto. Esta solución multitarea ha mejorado el rendimiento inicial significativamente, llegando a reducir un 300% el tiempo total de ejecución.

5.6. PUBLICACIÓN DE RESULTADOS

Por último, se ha desarrollado un proceso externo que permite publicar los resultados obtenidos en un servicio web para poder ser utilizados por un cliente. En este caso, la AEMPS se encarga de recoger estos datos que se encuentran en un formato acordado entre ambas partes.

Para la visualización de los resultados obtenidos por el sistema se genera un fichero XML por cada ficha técnica introducida según las especificaciones de la AEMPS. En la Figura 10 se muestra el esquema del fichero XML en formato XSD. El archivo tiene una estructura en la que se indica el código de la ficha técnica en un atributo y la información se recoge por secciones en elementos de tipo *Seccion*, permitiendo la escalabilidad en futuros trabajos. En los ficheros XML solo se muestran las secciones 1 (nombre del medicamento) y 4.1 (indicaciones terapéuticas).

Cada sección cuenta con un elemento de tipo *texto* en el que se recoge la información original y otro de tipo *SNOMED* para los conceptos SNOMED-CT encontrados. El trabajo de reconocimiento de enfermedades y mapeo con la terminología SNOMED-CT se ha realizado solo con la sección 4.1 por lo que solo esa sección cuenta con el elemento de tipo *SNOMED.* Por cada enfermedad identificada se crea un elemento de tipo *Enfermedad.* En cada elemento se recogen los conceptos candidatos de la terminología SNOMED-CT encontrados para cada enfermedad. Estos candidatos se recogen en elementos de tipo *Concepto_SNOMED* dejando en primer lugar el candidato favorito.

De forma paralela, se ha creado un servicio web basado en REST que permite la publicación de los resultados. REST es una arquitectura centrada en recursos que permite mayor simplicidad en los procesos de comunicación y que utiliza URIs como identificadores únicos de cada recurso. El servicio web utiliza y devuelve los ficheros XML resultantes para su publicación. El servicio funciona mediante la URI y el identificador de la ficha técnica. La petición puede devolver la información tanto en JSON como en XML. En la Tabla 10 se muestran las diferentes llamadas

que se pueden hacer al servicio. Actualmente, las secciones que cubre la API son la 1 (Nombre del medicamento) y la 4.1 (Indicaciones terapéuticas).

```
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema" elementFormDefault="qualified">
  <xs:element name="FT">
    <xs:complexType>
      <xs:sequence>
        <xs:element maxOccurs="unbounded" ref="Seccion"/>
      </xs:sequence>
      <xs:attribute name="id" use="required" type="xs:NCName"/>
    </xs:complexType>
  </xs:element>

  <xs:element name="Seccion">
    <xs:complexType>
      <xs:sequence>
        <xs:element minOccurs="0"  ref="SNOMED"/>
        <xs:element ref="texto"/>
      </xs:sequence>
      <xs:attribute name="id" use="required" type="xs:decimal"/>
    </xs:complexType>
  </xs:element>

  <xs:element name="SNOMED">
    <xs:complexType>
      <xs:sequence>
        <xs:element maxOccurs="unbounded" ref="Enfermedad"/>
      </xs:sequence>
    </xs:complexType>
  </xs:element>

  <xs:element name="Enfermedad">
    <xs:complexType>
      <xs:sequence>
        <xs:element ref="Concepto_SNOMED" minOccurs="0" maxOccurs="3" />
      </xs:sequence>
    </xs:complexType>
  </xs:element>

  <xs:element name="Concepto_SNOMED" type="xs:string"/>
  <xs:element name="texto" type="xs:string"/>
</xs:schema>
```

Figura 10. Esquema del fichero XML en XSD

Recurso	Descripción
GET FT/:id	Devuelve la ficha técnica especificada por el identificador (ej. F00036) con sus secciones y los conceptos de la terminología SNOMED-CT identificados en la misma.
GET FT/:id/seccion/:idSeccion	Devuelve la sección correspondiente al identificador de la ficha técnica especificada.
GET FT/:id/seccion/:idSeccion/SNOMED	Devuelve los conceptos SNOMED-CT identificados en la sección especificada de la ficha técnica.

Tabla 10. API REST

Capítulo 6. EVALUACIÓN

Con los resultados obtenidos para el conjunto inicial de 1078 fichas técnicas se ha procedido a la evaluación de la instanciación del método. El proceso de evaluación se focaliza en los resultados de las dos principales fases del método de forma aislada. La primera evaluación se centra en los resultados obtenidos en la fase de reconocimiento de enfermedades en las indicaciones terapéuticas. La segunda evaluación se centra en los resultados obtenidos en la fase de mapeo de la enfermedad con la terminología SNOMED-CT. La segunda evaluación se ha realizado mediante un *gold standard* proporcionado por los expertos de la AEMPS.

6.1. RESULTADOS DE LA FASE DE RECONOCIMIENTO DE ENFERMEDADES

Para la evaluación de las enfermedades detectadas por parte del sistema se ha extraído un subconjunto de 50 fichas técnicas de manera aleatoria y se han identificado manualmente las enfermedades que aparecen en las indicaciones terapéuticas para la creación de un *gold standard*. El *gold standard* se ha creado sin los expertos de la AEMPS. La Tabla 11 muestra los resultados de los experimentos realizados en el proceso de identificación de enfermedades para esas 50 fichas. Las métricas que se han utilizado han sido la precisión y la exhaustividad. La precisión mide la fracción de enfermedades recuperadas que son relevantes y la exhaustividad mide la fracción de enfermedades recuperadas del total. Las enfermedades relevantes son todas aquellas se consideran ciertas dentro del corpus y las enfermedades recuperadas aquellas que el sistema ha encontrado. La intersección entra ambas constituyen el grupo de enfermedades que el sistema ha encontrado y son correctas (verdaderos positivos).

$$Precisión = \frac{\{\text{enfermedades relevantes}\} \cap \{\text{enfermedades recuperadas}\}}{\{\text{enfermedades recuperadas}\}}$$

$$Exhaustividad = \frac{\{\text{enfermedades relevantes}\} \cap \{\text{enfermedades recuperadas}\}}{\{\text{enfermedades relevantes}\}}$$

Se ha estudiado por separado los resultados obtenidos utilizando los patrones y el diccionario de enfermedades MedDRA y los resultados obtenidos únicamente por los patrones. El subconjunto de fichas técnicas contiene un total de 152 enfermedades. En el primer caso, se han recuperado 147 correctas y 2 incorrectas. Se ha obtenido una precisión del 98.6% y una exhaustividad del 96%. Los resultados del sistema de reconocimiento de enfermedades basado solamente en patrones empeoran respecto el total. La precisión es del 100% pero la

exhaustividad baja hasta un 82%. Los resultados demuestran la hipótesis H1 de que el apoyo de un diccionario especializado en el dominio consigue aumentar en gran medida la recuperación de aquellas enfermedades que no se encuentran mediante patrones léxico-sintácticos, a pesar de que puedan añadir un pequeño porcentaje de falsos positivos.

Tipo de sistema de reconocimiento de enfermedades	Enfermedades recuperadas	Precisión	Exhaustividad
Sistema de reconocimiento de enfermedades basado en patrones y diccionario	149 - Correctas: 147 - Incorrectas: 2	147/149 (98.6 %)	147/152 (96 %)
Sistema de reconocimiento de enfermedades basado en patrones	125 - Correctas: 125 - Incorrectas: 0	125/125 (100 %)	125/152 (82 %)

Tabla 11. Evaluación del reconocimiento de enfermedades

Otro dato destacable es que la precisión que ejercen los patrones léxico-sintácticos sobre el sistema es el resultado de un estudio del corpus exhaustivo junto con expertos de la AEMPS. Para cada patrón se han detectado casi todas las excepciones que inducen a error (ej., *tratamiento de adultos*). Las reglas producidas contienen dicha información y la búsqueda sobre el corpus no produce falsos positivos.

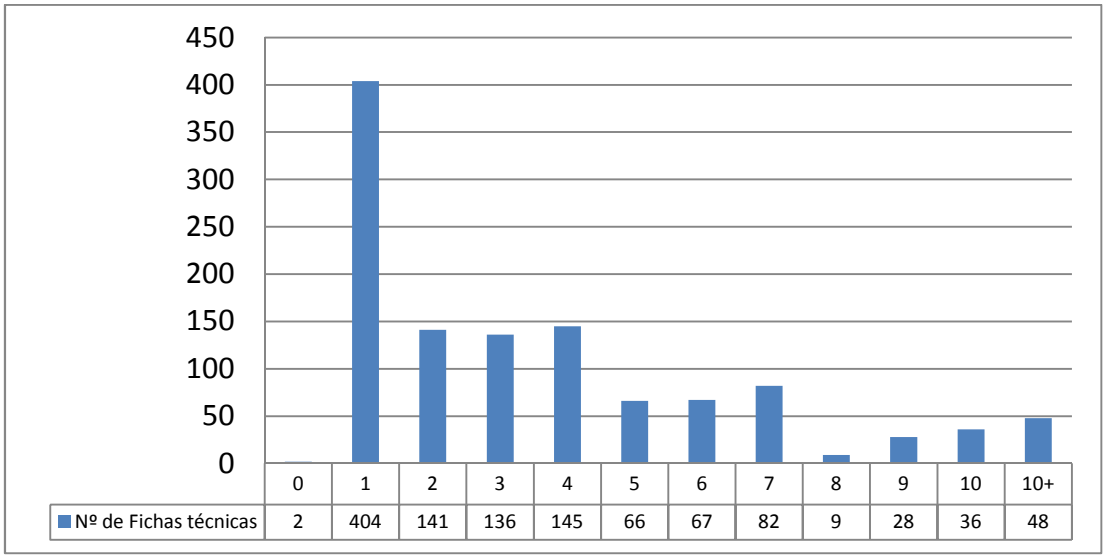


Figura 11. Histograma de enfermedades detectadas en fichas técnicas

Por último, se ha realizado un histograma reflejado en la Figura 11 en el que se muestra la frecuencia de enfermedades detectadas por ficha técnica. El histograma corresponde al sistema de reconocimiento de enfermedades basado en patrones y diccionario y está realizado sobre el corpus completo entregado por la AEMPS. Este histograma se ha realizado repetidas veces durante la instanciación del método para estudiar, junto con los expertos de la AEMPS, los resultados extremos

(cero enfermedades detectadas y diez o más enfermedades detectadas), para la detección de fallos en los patrones y su refinamiento. En los procesos de refinamiento se han encontrado nuevos patrones o se han solucionado fallos en la ejecución de las reglas sobre el corpus. Finalmente, el histograma muestra que en la mayoría de las indicaciones terapéuticas se detecta una sola enfermedad y que pocas superan más de diez. Las dos indicaciones terapéuticas en las que no se detectan enfermedades son fichas técnicas de anestesias generales.

6.2. RESULTADOS DE LA FASE DE MAPEO CON SNOMED-CT

La evaluación de los resultados obtenidos en la fase de mapeo con la terminología SNOMED-CT se ha realizado mediante un *gold standard* proporcionado por los expertos del dominio. La AEMPS ha realizado un proceso de evaluación aleatorio de 20 fichas técnicas indicando si el concepto o conceptos SNOMED-CT mapeados son correctos o erróneos. Las fichas técnicas del *gold standard* de la AEMPS son diferentes a las del *gold standard* creado para la evaluación de las enfermedades detectadas.

El resultado de esta evaluación se puede ver en la Tabla 12. La primera fila corresponde a la evaluación del sistema de mapeo completo. Del total de 67 enfermedades identificadas y mapeadas con la terminología SNOMED-CT, 58 se han validado como correctas y 9 como erróneas. Como en el apartado anterior, se hace un cálculo de la precisión y exhaustividad del sistema, obteniendo un 86,5% en ambas. La igualdad entre la precisión y exhaustividad se debe a que el número de falsos positivos obtenidos por el sistema es el mismo que los conceptos SNOMED-CT correctos que tendría que devolver el sistema (falsos negativos).

Para demostrar la hipótesis H3 de que la generación de combinaciones de las enfermedades mediante el número de las palabras y los sinónimos aumenta la precisión de los mapeos, se ha vuelto a evaluar los resultados de los mapeos de las enfermedades sin la generación de combinaciones. El resultado se puede visualizar en la segunda fila de la Tabla 12. La precisión y la exhaustividad disminuyen a un 73% por lo que demostramos que la generación de combinaciones para mapeos basados en comparaciones exactas de cadenas aumenta la precisión.

De la misma manera, se ha demostrado la hipótesis H2 de que los moderadores lingüísticos usados como puntos de corte en las reglas sirven para evitar ruido en el proceso de mapeo. La tercera fila de la Tabla 12 muestra, para las mismas enfermedades, los resultados obtenidos en el proceso de mapeo sin los moderadores lingüísticos. La precisión y exhaustividad disminuye a un 70%.

Tipo de mapeos realizados para las 20 FT	Enfermedades mapeadas	Precisión	Exhaustividad
El sistema completo	67 enfermedades - Correctas: 58 - Incorrectas: 9	58/67 (86,5 %)	58/67 (86,5 %)
Sin generación de combinaciones de la enfermedad	67 enfermedades -Correctas 49 -Incorrectas 18	49/67 (73%)	49/67 (73%)
Sin moderadores lingüísticos	67 enfermedades - Correctas: 47 - Incorrectas: 20	47/67 (70%)	47/67 (70%)
Sin generación de combinaciones de la enfermedad ni moderadores lingüísticos	67 enfermedades - Correctas: 38 - Incorrectas: 29	38/67 (56%)	38/67 (56%)

Tabla 12. Evaluación con el *gold standard* proporcionado por la AEMPS

El último caso de estudio, es en el que aplicamos el mismo proceso de mapeo sin generar combinaciones y sin usar moderadores lingüísticos. Este es el peor de los casos, ya que la precisión del algoritmo de mapeo disminuye a un 56%.

6.3. DISCUSIÓN

La complejidad del proyecto realizado ha consistido principalmente en el reconocimiento de enfermedades sobre las indicaciones terapéuticas. Los resultados obtenidos han sido el producto del estudio exhaustivo del corpus en busca de patrones léxico-sintácticos y de palabras que permitan acotar la enfermedad detectada para evitar ruido en el sistema. La calidad de los resultados que se logren durante este proceso es directamente proporcional al resultado final del proceso de mapeo con la terminología SNOMED-CT. En la sección 6.1 se ha mostrado la precisión y exhaustividad con la que trabajan los patrones. El porcentaje que añade el uso del diccionario especializado como MedDRA supone una pieza clave en la detección de enfermedades ya que identifica casi a la totalidad de aquellas enfermedades que están fuera del alcance de los patrones. La mayoría de los falsos negativos que nos encontramos por las reglas léxico-sintácticas en el proceso son conceptos que se van a repetir varias veces sobre la misma indicación terapéutica con otros patrones. No obstante, los conceptos sueltos están ligados directamente al diccionario y permiten completar la información relevante de la ficha técnica.

La evaluación del algoritmo de mapeo con la terminología SNOMED-CT es insuficiente debido al bajo número de fichas técnicas en el *gold standard*. Por ello, se ha decidido a realizar un estudio completo de los resultados obtenidos por el proceso. Se ha querido mostrar la utilidad real que el sistema puede llegar a tener, ya que el mapeo erróneo de una enfermedad cuyo índice de aparición es muy bajo

no debería contar igual que una con un índice de aparición mayor. Del total de las 1078 fichas técnicas proporcionadas por la AEMPS se han estudiado todos los resultados obtenidos y su número de apariciones. El resultado se refleja en la Figura 12. Del total de 4553 enfermedades detectadas y mapeadas, 3126 han sido resueltas (69 %) y 107 han fallado (2%). El proceso de verificación se ha realizado mediante los comentarios recibidos por parte de la AEMPS y por la similitud exacta entre la enfermedad y el concepto SNOMED-CT. 1320 enfermedades se han marcado como duda (29%); este grupo involucra a los conceptos resultantes en los que se necesitaría un experto en el dominio para su verificación. Los problemas que no permiten la comprobación del resultado son:

- *Problemas de granularidad*. Este problema aparece cuando se introduce una enfermedad y el sistema devuelve una menos específica o sin tanto nivel de gradación (ej., Para *leucemia meníngea* el sistema devuelve *leucemia*). Este tipo de resultados se deben al algoritmo de búsqueda utilizado. Para no devolver falsos positivos, se devuelve un concepto más general basado en las palabras clave en vez de hacer una aproximación por todas sus palabras. Se necesita una revisión por parte de expertos para determinar si estos resultados son validos o no. Este problema constituye un 18% del total de los resultados obtenidos.

- *Términos desconocidos en SNOMED-CT*. Estos términos corresponden a enfermedades para las que no se ha encontrado un concepto similar ni aproximado dentro de la terminología. No se clasifican dentro del grupo de error ya que son conceptos generales que no se reflejan en la terminología SNOMED-CT (ej., *morbilidad*). Se necesita a un experto del dominio para determinar si estas enfermedades siguen siendo relevantes. Este problema aparece el 5% del total de las veces.

- *Múltiples conceptos iguales en SNOMED-CT*. Incluso con la terminología SNOMED-CT podada, existen conceptos que se repiten dentro de grupos de enfermedades. A veces dentro del mismo grupo (ej., existen tres conceptos con el nombre *neoplasia* dentro del grupo de *trastornos*). El problema aparece un 3% del total de las veces.

- *Conceptos SNOMED-CT fuera de alcance del dominio*. Existen fichas técnicas que describen anestésicos o anticonceptivos. Para la AEMPS esta información es relevante, pero el sistema no es capaz de devolver el concepto debido a que durante la extracción del subconjunto de enfermedades de la terminología SNOMED-CT se eliminan sustancias y productos. El problema aparece un 2% del total de las veces.

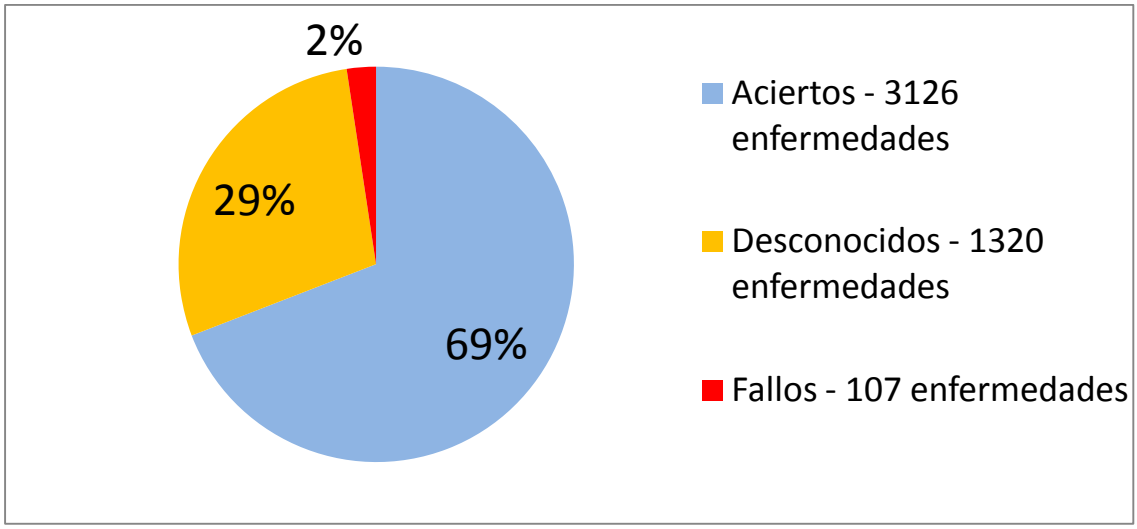


Figura 12. Tasa de aciertos sobre el total de enfermedades mapeadas

Capítulo 7. CONCLUSIONES

Las contribuciones principales de esta tesis son:

1. Una nueva técnica para el reconocimiento de entidades nombradas de enfermedades basado en reglas léxico-sintácticas y *gazetteers* especializados en el dominio médico.
2. Un nuevo algoritmo de mapeo basado combinaciones y en las funciones sintácticas de las palabras.
3. Un método para la utilización de la terminología SNOMED-CT en el proceso de reconocimiento y mapeo de enfermedades en fichas técnicas de medicamentos.
4. La instanciación del método de reconocimiento de entidades nombradas y de mapeo con la terminología SNOMED-CT para un caso de uso real de la AEMPS.

La técnica de reconocimiento de entidades propuesta en esta tesis podría ser reutilizada para cualquier dominio. Las reglas léxico-sintácticas son una de las principales técnicas que se usan en las tareas de identificación de entidades nombradas. La combinación de estas reglas junto a los *gazetteer* de moderadores lingüísticos permite acotar las entidades nombradas identificadas y evitar que se introduzca ruido. La calidad de los resultados obtenidos es debida a un estudio meticuloso del corpus y de una generación manual completa de las reglas con todas sus posibles excepciones.

Para evitar que la identificación de entidades se quede corta, el método de búsqueda se apoya en un diccionario terminológico para identificar aquellas enfermedades que no aparecen junto a patrones léxico-sintácticos.

El algoritmo de mapeo terminológico propuesto explota las funciones sintácticas que las palabras desempeñan en la frase. Los nombres actúan como palabras clave y el resto como normales. La diferenciación de las palabras clave de las que no lo son permite obtener resultados más aproximados en caso de que no se encuentre un mapeo directo. Esta aproximación permite reducir la aparición de falsos positivos en gran medida ya que en el ámbito medico tienden a minimizarse lo máximo posible. Los conceptos que no consiguen mapearse de manera directa acaban haciéndolo con otro concepto más general de la terminología.

Las técnicas léxicas de mapeo basadas en comparaciones de cadenas mostradas en el estado del arte en la sección 2.2 no son útiles con conceptos extraídos de textos en lenguaje natural en español. La estructura de la oración del español es menos restrictiva que la de otros idiomas por lo que la colocación de adjetivos y nombres no siempre están en un orden establecido. Además, los adjetivos en inglés son

invariantes en género y en número mientras que en español no. Estos dos problemas hacen que la metodología descarte las técnicas de comparación de cadenas excepto la comparación exacta de palabras. Dos cadenas con conceptos idénticos donde aparezca alguno de los problemas mencionados dan porcentajes de similitud bajos. Junto a estos problemas, también hay que incluir las palabras con tilde que no aparecen en las terminologías tras el proceso de traducción.

Actualmente existen pocos trabajos que expongan casos prácticos para la terminología SNOMED-CT. La terminología SNOMED-CT es muy grande y compleja de manejar. El proceso de inferir relaciones o información adicional no es trivial. Los códigos asociados a cada concepto no aportan ningún tipo de información por lo que siempre es necesario acceder a las relaciones de cada concepto para poder ser explotadas.

Incluso con un algoritmo de mapeo basado en palabras clave para dar resultados más concretos, se ha podado la terminología SNOMED-CT para facilitar el procesamiento y evitar conceptos fuera del dominio de las enfermedades. La terminología SNOMED-CT cubre una gran cantidad de definiciones y términos que se aplican en diferentes ámbitos.

Capítulo 8. LÍNEAS FUTURAS

Las líneas futuras de la tesis están orientadas a aumentar la base de conocimientos que el método utiliza en sus procesos. Parte de la fase de generación de recursos de conocimiento podría delegarse en otros sistemas externos que ayudasen a inferir conocimiento básico. De esta manera se ahorraría tiempo en la ejecución de las tareas manuales del estudio del corpus.

Una de esas tareas es la identificación de sinónimos, acrónimos y siglas. Esta tarea del proceso de estudio del corpus involucra bastante esfuerzo por parte de revisores, ya sean expertos o no. Muchos de los sinónimos que han sido identificados por humanos podrían haberse obtenido de manera automática mediante un diccionario relacionado o una ontología. Por ejemplo, los conceptos *cáncer* y *déficit* tienen como sinónimos *tumor maligno* y *carencia,* respectivamente. Ambos se usan en el ámbito médico pero su relación semántica no es tan concreta como para que solo pudiese detectarla un experto.

Una línea futura se focaliza en la fase de identificación de entidades nombradas. La creación de más patrones junto con mejores diccionarios que apoyen la búsqueda haría que las enfermedades detectadas aumentasen y fuesen más robustas frente a cambios del lenguaje natural. La creación de reglas identificando excepciones y posibles antipatrones es un proceso costoso en cuanto a tiempo y mantenimiento.

Incluso una vez creadas, estas reglas siguen estando ligadas al dominio y su cambio implica una nueva revisión y creación de las mismas reglas. Una línea futura sería hacer posible la identificación de excepciones mediante anotaciones independientes que permitiese agilizar el proceso de creación de las reglas basadas en patrones léxico-sintácticos.

Otra línea futura fuera de los recursos de conocimiento se focaliza en la utilización de otras terminologías para el proceso de mapeo. Aunque SNOMED-CT es una de las más completas que hay actualmente, existen otras terminologías de dominios más específicos o con mayor relevancia en la práctica clínica como UMLS.

El trabajo realizado en esta tesis podría extenderse a otras secciones de las fichas técnicas de medicamentos en las que puede ser relevante identificar enfermedades y mapearlas (ej., la sección de reacciones adversas). Además, otros trabajos futuros podrían utilizar los métodos propuestos en esta tesis para otro idioma, ya que el trabajo realizado se encuentra en español así como los recursos que se utilizan.

APÉNDICE A

Palabras con más de 80 apariciones en el corpus proporcionado por la AEMPS.

Palabra	Frecuencia	Palabra	Frecuencia
TRATAMIENTO	2130	DEPRESIVOS	132
PACIENTES	984	SÍNTOMAS	127
ADULTOS	701	EDAD	122
INDICADO	647	SANDOZ	118
ENFERMEDAD	433	TEVA	118
TRASTORNO	412	REFLUJO	117
PROLONGADA	365	CANDESARTÁN	115
VER	356	LEVE	115
SECCIÓN	306	ARTRITIS	113
HIPERTENSIÓN	276	METOJECT	110
PREVENCIÓN	270	PEN	110
MAYORES	228	PHARMACIA	110
AÑOS	224	AUROBINDO	109
ARTERIAL	221	USO	104
COMBINACIÓN	217	ANSIEDAD	103
EPISODIOS	200	SOCIAL	101
SINTOMÁTICO	196	TIPO	101
GRAVE	190	ALZHEIMER	99
INDICADA	185	ASOCIACIÓN	99
ESENCIAL	181	ÚLCERAS	99
PUEDE	181	OLANZAPINA	98
NIÑOS	179	PHARMA	96
DOLOR	178	EFECTO	95
BIPOLAR	174	FIJAS	95
MONOTERAPIA	174	CONTROLARSE	93
ADECUADAMENTE	173	ORAL	92
RIESGO	171	CINFA	91
ADOLESCENTES	164	IDIOPÁTICA	91
CARDIOVASCULAR	164	INSUFICIENCIA	90
INFECCIONES	162	HA	89
MODERADA	161	DURANTE	88
PRESIÓN	161	GRAVES	88
CUANDO	159	SANOFI	88
HAN	154	INCLUYENDO	87
ES	153	SIDO	87
CUYA	152	VALSARTÁN	87

TERAPIA	147	**STADA**	85
DEBE	136	**JERINGA**	84
FLUCTUACIONES	136	**DESPUÉS**	83
PERFUSIÓN	136	**REDUCCIÓN**	82

APÉNDICE B

Ejemplos de patrones léxico-sintácticos identificados en el corpus proporcionado por la AEMPS.

Patrones	Ejemplo
Patrón Accidentes	*Accidentes cardiovasculares*
Patrón Alivio	Alivio del *dolor muscular*
Patrón Alivio Sintomático	Alivio sintomático de la *sequedad ocular*
Patrón Algia	*Polimialgia reumática*
Patrón Angina	*Angina de pecho estable*
Patrón Antecedentes De	Antecedentes de *cardiopatía coronaria*
Patrón Anticoncepción	*Anticoncepción oral*
Patrón Asis	*Candidiasis vulvovaginal*
Patrón Asociadas Con	Asociadas con *producción excesiva de mucosa*
Patrón Asociados Con	Asociados con *trastorno bipolar*
Patrón Cáncer	*Cáncer gástrico*
Patrón Crisis	*Crisis mioclónicas*
Patrón Curada	*Esofagitis curada*
Patrón Diabetes	*Diabetes tipo 2*
Patrón Diagnostico De	Diagnostico de *epilepsia*
Patrón Disfunción	*Disfunción eréctil*
Patrón Dolor	*Dolor postquirúrgico*
Patrón Edema	*Edema angioneurótico*
Patrón Ema	*Eritema severo multiforme*
Patrón Emia	*Bacteriemia*
Patrón Enfermedad	*Enfermedad por reflujo gastroesofágico*
Patrón Enfermedades	*Enfermedades gastrointestinales*
Patrón Episodios	*Episodios depresivos mayores*
Patrón Episodios Recurrentes	Episodios recurrentes del *herpes genital*
Patrón Estados	*Estados carenciales del hierro*
Patrón Exacerbación	*Exacerbación aguda de la bronquitis crónica*
Patrón Fiebre	*Fiebre tifoidea*
Patrón Formas Graves	Formas graves del *acné*
Patrón Gripe	*Gripe A*
Patrón Hemorragias	*Hemorragias intracerebrales*
Patrón Hiper	*Hipertiroidismo*
Patrón Hipo	*Hipocalcemia*
Patrón Indicado En	Indicado en la *profilaxis del asma*
Patrón Infección	*Infección de la vesícula biliar*
Patrón Infecciones	*Infecciones graves del tracto urinario*

Patrón Infectados	Infectados con *VIH*
Patrón Inmunización	Inmunización activa frente a *tétanos*
Patrón Insuficiencia	*Insuficiencia cardiaca crónica*
Patrón Itis	*Artritis reumatoide*
Patrón Leucemia	*Leucemia linfocítica crónica*
Patrón Linfoma	*Linfomas no hodgkinianos*
Patrón Lupus	*Lupus eritematoso*
Patrón Moderado a grave	*Acné* de moderado a grave
Patrón Neumonía	*Neumonía nosocomial*
Patrón Obstrucción	*Obstrucción nasal*
Patrón Oma	*Sarcoma osteógeno*
Patrón Orrea	*Dismenorrea*
Patrón Osis	*Osteoporosis*
Patrón Pacientes Con	Pacientes con *glaucoma de ángulo abierto*
Patrón Patología	*Patología ocular*
Patrón Penfigo	*Pénfigo penfigoide bulloso*
Patrón Penia	*Trobocitopenia idiopática*
Patrón Plasia	*Hiperplasia benigna de próstata*
Patrón Plastia	*Angioplastia coronaria*
Patrón Prevención	*Prevención de eventos cardiovasculares*
Patrón Procesos	*Procesos catarrales y gripales*
Patrón Profilaxis	*Profilaxis del rechazo agudo*
Patrón Reducción De	Reducción de la *morbilidad cardiovascular*
Patrón Riesgo De	Riesgo de *fractura ósea*
Patrón Síntomas De	Síntomas de *abstinencia de nicotina*
Patrón Sufren	Pacientes que sufren *angina de pecho estable*
Patrón Tétanos	*Tétanos*
Patrón Tos	*Tos ferina*
Patrón Trastornos	*Trastorno de angustia*
Patrón Tratamiento	Tratamiento adyuvante en *cáncer de mama avanzado*
Patrón Tratamiento De	Tratamiento de *náuseas y vómitos*
Patrón Tumores	*Tumores tiroideos malignos*
Patrón Urticaria	*Urticaria*
Patrón Virus	*Virus de la Inmunodeficiencia Humana*
Patrón Vacunación	Vacunación frente a *varicela*

BIBLIOGRAFÍA

Abacha, A. B., & Zweigenbaum, P. (2011). Medical entity recognition: a comparison of semantic and statistical methods. In *2011 Workshop on Biomedical Natural Language Processing* (pp. 56–64).

Allones, J. L. I., Hernández, D. M., & Taboada, M. (2014). Automated Mapping of Clinical Terms into SNOMED-CT. An Application to Codify Procedures in Pathology. *J. Medical Systems 38*, 110–120.

Aronson, A. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *AMIA Symposium*, 17–21.

Baluja, S., Mittal, V. O., & Sukthankar, R. (2000). Applying Machine Learning for High-Performance Named-Entity Extraction. *Computational Intelligence*, *16*, 586–595.

Batool, R., Khattak, A. M., Kim, T.-S., & Lee, S. (2013). Automatic extraction and mapping of discharge summary's concepts into SNOMED CT. *Conference Proceedings: Annual International Conference Of The IEEE Engineering In Medicine And Biology Society.*, *2013*, 4195–4198.

Bodenreider, O. (2008). Issues in mapping LOINC laboratory tests to SNOMED CT. *AMIA Annual Symposium Proceedings*, 51–55.

Bodenreider, O., & Zhang, S. (2006). Comparing the representation of anatomy in the FMA and SNOMED CT. *AMIA Annual Symposium Proceedings*, 46–50.

Bodnari, A., Del, L., & Lavergne, T. (2013). A Supervised Named-Entity Extraction System for Medical Text. *Proceedings of the ShARe/CLEF Evaluation Lab*, 1–8.

Cao, F., Sun, X., Wang, X., Li, B., Li, J., & Pan, Y. (2011). Ontology-based knowledge management for personalized adverse drug events detection. In *Studies in Health Technology and Informatics* (Vol. 169, pp. 699–703).

Castro, E., Iglesias, A., Martínez, P., & Castaño, L. (2010). Automatic identification of biomedical concepts in spanish-language unstructured clinical texts. *Proceedings of the 1st ACM International Health Informatics Symposium*, 751–757.

Castro, E., & Martinez, P. (2007). Evaluation of a Named-Entity Recognition System over SNOMED CT.

Choi, N., Song, I.-Y., & Han, H. (2006). A survey on ontology mapping. *ACM SIGMOD Record*, *35*, 34–41.

Cruanes, J., Romá-Ferri, M. T., & Lloret, E. (2012). Measuring lexical similarity methods for textual mapping in nursing diagnoses in Spanish and SNOMED-CT. *Studies in Health Technology and Informatics*, *180*, 255–259.

De Silva, T. S., MacDonald, D., Paterson, G., Sikdar, K. C., & Cochrane, B. (2011). Systematized nomenclature of medicine clinical terms (SNOMED CT) to represent computed tomography procedures. *Computer Methods and Programs in Biomedicine*, *101*, 324–329.

Elkin, P. L., Brown, S. H., Husser, C. S., Bauer, B. A., Wahner-Roedler, D., Rosenbloom, S. T., & Speroff, T. (2006). Evaluation of the content coverage of SNOMED CT: ability of SNOMED clinical terms to represent clinical problem lists. *Mayo Clinic Proceedings. Mayo Clinic*, *81*, 741–748.

Euzenat, J., & Shvaiko, P. (2007). *Ontology matching*. *Heidelberg: Springer.* doi:10.1007/978-3-540-49612-0

Fung, K. W., & Bodenreider, O. (2005). Utilizing the UMLS for semantic mapping between terminologies. *AMIA Annual Symposium Proceedings / AMIA Symposium. AMIA Symposium*, 266–270.

Garde, S., Knaup, P., Hovenga, E., & Heard, S. (2007). Towards semantic interoperability for electronic health records. *Methods of Information in Medicine*, *46*, 332–343.

Giannangelo, K., & Millar, J. (2012). Mapping SNOMED CT to ICD-10. *Studies in Health Technology and Informatics*, *180*, 83–87.

Han, X., & Ruonan, R. (2011). The Method of Medical Named Entity Recognition Based on Semantic Model and Improved SVM-KNN Algorithm. *2011 Seventh International Conference on Semantics, Knowledge and Grids*, 21–27.

Hina, S., Atwell, E., & Johnson, O. (2010). Secure information extraction from clinical documents using SNOMED CT gazetteer and natural language processing. *2010 International Conference for Internet Technology and Secured Transactions, ICITST 2010*, 1–5.

Ikonomakis, M. (2005). Text classification using machine learning techniques. *WSEAS Transactions on Computers*, *4*(8), 966–974.

Jain, A. K., Jain, A. K., Duin, R. P. W., Duin, R. P. W., Mao, J., & Mao, J. (2000). Statistical Pattern Recognition: A Review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*, 4–37.

James, A. G., & Spackman, K. A. (2008). Representation of disorders of the newborn infant by SNOMED CT. *Studies in Health Technology and Informatics*, *136*, 833–838.

Jiang, G., & Chute, C. G. (2009). Auditing the Semantic Completeness of SNOMED CT Using Formal Concept Analysis. *Journal of the American Medical Informatics Association : JAMIA*, *16*, 89–102.

Kashyap, V., & Sheth, A. (1996). Semantic and schematic similarities between database objects: a context-based approach. *The VLDB Journal The International Journal on Very Large Data Bases*, *5*, 276–304.

Kazama, J., & Torisawa, K. (2007). Exploiting Wikipedia as External Knowledge for Named Entity Recognition. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (pp. 698–707).

Kim, H.-Y., & Park, H.-A. (2012). Development and evaluation of data entry templates based on the entity-attribute-value model for clinical decision support of pressure ulcer wound management. *International Journal of Medical Informatics*, *81*, 485–492.

Lee, D., de Keizer, N., Lau, F., & Cornet, R. (2014). Literature review of SNOMED CT use. *Journal of the American Medical Informatics Association : JAMIA*, *21*, 1–9.

Lee, N.-J., & Bakken, S. (2007). Development of a prototype personal digital assistant-decision support system for the management of adult obesity. *International Journal of Medical Informatics*, *76 Suppl 2*, S281–S292.

Lei Zeng, M., & Mai Chan, L. (2004). Trends and issues in establishing interoperability among knowledge organization systems. *Journal of the American Society for Information Science and Technology*, *55*, 377–395.

Maynard, D., Tablan, V., & Ursu, C. (2001). Named entity recognition from diverse text types. *Recent Advances in Natural Language Processing*, 257–274.

Meizoso García, M., Iglesias Allones, J. L., Martínez Hernández, D., & Taboada Iglesias, M. J. (2012). Semantic similarity-based alignment between clinical archetypes and SNOMED CT: An application to observations. *International Journal of Medical Informatics*, *81*, 566–578.

Meizoso, M., Allones, J. L., Taboada, M., Martinez, D., & Tellado, S. (2011). Automated mapping of observation archetypes to SNOMED CT concepts. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 6686 LNCS, pp. 550–561).

Nadeau, D., & Sekine, S. (2006). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, *30*, 3–26.

Nadeau, D., Turney, P. D., & Matwin, S. (2006). Unsupervised named-entity recognition: Generating gazetteers and resolving ambiguity. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 4013 LNAI, pp. 266–277).

Oronoz, M., Casillas, A., Gojenola, K., & Perez, A. (2013). Automatic Annotation of Medical Records in Spanish with Disease , Drug, 536–543.

Patrick, J., Wang, Y., & Budd, P. (2007). An automated system for conversion of clinical notes into SNOMED clinical terminology. *Conferences in Research and Practice in Information Technology Series*, *68*, 219–226.

Richman, A. E., & Schone, P. (2008). Mining Wiki Resources for Multilingual Named Entity Recognition. *Proceedings of ACL-08: HLT*, 1–9.

Ruch, P., Gobeill, J., Lovis, C., & Geissbühler, A. (2008). Automatic medical encoding with SNOMED categories. *BMC Medical Informatics and Decision Making*, *8 Suppl 1*, S6.

Shortliffe, E. H., & Cimino, J. J. (2006). *Biomedical Informatics*. *JAMA* (Vol. 296).

Shvaiko, P., & Euzenat, J. (2013). Ontology Matching: State of the Art and Future Challenges. *IEEE Transactions on Knowledge and Data Engineering*, *25*, 158–176.

Skeppstedt, M., Kvist, M., & Dalianis, H. (2007). Rule-based Entity Recognition and Coverage of SNOMED CT in Swedish Clinical Text. *Eighth International Conference on Language Resources and Evaluation, LREC 2012*, 1250–1257.

Snow, R., Jurafsky, D., & Ng, A. Y. (2005). Learning syntactic patterns for automatic hypernym discovery. *Nips*, *17*, 1297–1304.

Stenzhorn, H., Pacheco, E. J., Nohama, P., & Schulz, S. (2009). Automatic mapping of clinical documentation to SNOMED CT. In *Studies in Health Technology and Informatics* (Vol. 150, pp. 228–232).

Sun, J. Y., & Sun, Y. (2006). A System for Automated Lexical Mapping. *Journal of the American Medical Informatics Association*, *13*, 334–343.

Toral, A., & Mu, R. (2006). A proposal to automatically build and maintain gazetteers for Named Entity Recognition by using Wikipedia. *Communications*, 56–61.

Walker, J., Pan, E., Johnston, D., Adler-Milstein, J., Bates, D. W., & Middleton, B. (2005). The value of health care information exchange and interoperability. *Health Affairs (Project Hope)*, *Suppl Web* .

Wang, Y., Halper, M., Min, H., Perl, Y., Chen, Y., & Spackman, K. A. (2007). Structural methodologies for auditing SNOMED. *Journal of Biomedical Informatics*, *40*, 561–581.

Wang, Y., Halper, M., Wei, D., Gu, H., Perl, Y., Xu, J., … Hripcsak, G. (2012). Auditing complex concepts of SNOMED using a refined hierarchical abstraction network. *Journal of Biomedical Informatics*, *45*, 1–14.

Wyatt, J. C., & Liu, J. L. Y. (2002). Basic concepts in medical informatics. *Journal of Epidemiology & Community Health*, *56*, 808–812.

Zhang, L., Pan, Y., & Zhang, T. (2004). Focused named entity recognition using machine learning. *Proceedings of the 27th Annual International Conference on Research and Development in Information Retrieval - SIGIR '04*, 281.

Zhang, S., & Elhadad, N. (2013). Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts. *Journal of Biomedical Informatics*, *46*(6), 1088–1098.